CEDVS-BENCH: SPEAK OR STAY SILENT? A BENCHMARK FOR CONTINUOUS EVENT DETECTION IN VIDEO STREAMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video assistants must recognize events while they unfold, decide when evidence justifies a report, and remain silent when a requested event never occurs. Retrospective localization and answers at selected query times only partially capture this requirement. We introduce **CEDVS-Bench**, a unified benchmark for continuous detection of language-defined events in video streams. Candidate descriptions are specified before observation; models make causal decisions every second, without access to future frames or event boundaries. The benchmark contains 2,840 videos, 15,983 positive event instances, and 66,158 absent candidates across short, medium, and long video settings. Our evaluation jointly measures onset localization, decision timeliness, processing-aware response availability, and rejection of absent events. Experiments on both window-based VLMs and native streaming models show that accurate localization does not necessarily lead to timely reporting, while semantically similar absent events remain difficult to reject, especially over longer observation horizons. CEDVS-Bench provides a unified testbed for accurate, timely, and reliable continuous event monitoring in video streams.

1 INTRODUCTION

Video understanding is increasingly moving from retrospective analysis toward continuous assistance. A practical video assistant should not wait until a video ends before deciding what happened; it should monitor events as observations arrive and respond at the right moment. For example, given the instruction "Notify me when someone opens the cabinet," approaching the cabinet or touching its handle should not trigger an alert. The system should respond only when the cabinet is actually opened, and should otherwise remain silent. This setting is important for real-time video assistance because both delayed responses and false alarms reduce the usefulness of the system.

Existing benchmarks only partially capture this capability. Temporal grounding and event understanding benchmarks evaluate whether a model can recognize or localize an event after the relevant video content has been observed (Gao et al., 2017; Lei et al., 2021; Liu et al., 2024; Chen et al., 2025). Recent streaming benchmarks introduce causal observations, online question answering, dialogue, and proactive responses (Lin et al., 2026; Huang et al., 2025; Yang et al., 2025; Niu et al., 2025). However, these settings generally do not require a model to continuously maintain a prespecified set of event interests throughout the entire stream. Such persistent monitoring introduces a different challenge: the model must verify that an event has truly occurred, report it soon after onset, and consistently reject plausible events that never occur. These objectives are tightly coupled—aggressive triggering may improve recall but increase false alarms, while conservative behavior may suppress false alarms at the cost of missed or late detections.

We introduce **CEDVS-Bench**, a benchmark for continuous detection of language-defined events in video streams. Before observation, the model receives a persistent set of event candidates and makes a causal decision every second using only the video observed so far. The benchmark includes both occurring and explicitly absent events, and evaluates not only whether an event is localized correctly, but also whether it is reported in time and whether absent events are reliably rejected. Our contributions are:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: Continuous event monitoring requires persistent event interests, timely triggering, and sustained rejection of absent events.

- we formulate **continuous language-defined event monitoring** as a benchmark task that requires persistent full-stream monitoring, timely triggering, and sustained silence;
- we introduce **CEDVS-Bench** and a causal per-second evaluation protocol that jointly measures onset localization, decision timeliness, processing-aware response availability, and negative-event rejection;
- we evaluate both window-based VLMs and native streaming models, and further analyze window-based models across event difficulty, video duration, and event duration.

2 Related Work

Offline event localization and understanding. Temporal grounding and event-level video understanding aim to identify *what* happens and *when* it happens in a video. Charades-STA and QVHighlights provide language-conditioned temporal localization, while E.T. Bench extends evaluation to event-level grounding, captioning, and reasoning (Gao et al., 2017; Lei et al., 2021; Liu et al., 2024). CG-Bench further studies evidence-grounded understanding in long videos (Chen et al., 2025). Open-vocabulary and open-set temporal localization extend these settings to novel or unknown action categories (Gupta et al., 2024; Bao et al., 2022). These tasks establish strong semantic and temporal foundations, but are primarily retrospective: correct localization after observing the relevant video does not require the model to decide when to report an event during an ongoing stream.

Online and streaming video understanding. Recent benchmarks move from offline video understanding toward causal processing of video streams. StreamingBench evaluates understanding at multiple time points, OVBench studies perception, memory, and reasoning over online video, and SVBench introduces temporally linked multi-turn dialogue (Lin et al., 2026; Huang et al., 2025; Yang et al., 2025). OVO-Bench further includes real-time understanding and forward active responding (Niu et al., 2025). More recent work moves closer to continuous proactive evaluation: SPOT-Bench evaluates timely responses across streaming detection, interaction, and intervention tasks, while PaSBench-Video focuses on causally timed safety warnings and false positives on no-risk videos (Chatterjee et al., 2026; Zhao et al., 2026). CEDVS-Bench instead targets persistent

Table 1: Benchmark comparison. ✓: explicit support; △: a subset of tasks or a related but different criterion; ×: outside the stated evaluation. C: causal input; L: language-defined event interests; A: active triggering; N: negative rejection; T: temporal evaluation. Scale units follow the cited papers.

Benchmark	Task	Videos / annotations	C	L	A	N	T
StreamingBench	QA / proactive	900 / 4,500 QA	✓	△	△	△	△
OVBench	Online QA	1463 / 7090 labels	✓	△	×	△	△
OVO-Bench	QA / active	644 / ∼2.8K labels	✓	△	△	△	✓
SVBench	Dialogue	1,353 / 49,979 QA	✓	△	×	×	△
OmniMMI	Interaction	1,121+ / 2,290 QA	✓	△	✓	△	✓
E.T. Bench	Event understanding	7K / 7.3K samples	×	✓	×	×	✓
SPOT-Bench	Perception / assistance	940 / task-specific	✓	✓	✓	△	✓
PaSBench-Video	Safety warnings	740 / risk, no-risk	✓	△	✓	✓	✓
CEDVS-Bench	**Event detection**	**2,840 / 15,983 inst.**	✓	✓	✓	✓	✓

N includes several rejection targets: premature output, hallucination, irrelevant input, or extra predictions during non-response periods. SPOT already penalizes full-stream over-response; PaS uses no-risk videos. Our negatives are pre-specified absent events, with semantic difficulty labels where available. T includes localization or response windows and does not necessarily include processing cost.

monitoring of a pre-specified set of language-defined events, jointly evaluating onset localization, timely triggering, and rejection of explicitly absent candidates throughout the stream.

Long-video and streaming-capable models. A separate line of work develops models that can process long or continuously arriving video. LongVA and LongVU extend effective video context, while VideoLLM-online and Dispider support native streaming or active real-time interaction (Zhang et al., 2024a; Shen et al., 2024; Chen et al., 2024; Qian et al., 2025). These advances improve how streaming video is processed, but do not by themselves define how persistent event monitoring should be evaluated. We therefore test both window-based VLMs and native streaming models under the same continuous event-detection protocol, explicitly measuring localization, reporting timeliness, and false triggering.

3 Task, Evaluation, and Dataset Construction

CEDVS-Bench evaluates continuous detection of language-defined events in streaming video. Before a video begins, the model is given a set of event descriptions to monitor. As the video arrives, it must continuously determine, using only observations available so far, whether any monitored event has occurred. We first formalize the continuous detection task, then introduce the evaluation protocol, and finally describe the unified construction pipeline and the three subsets of CEDVS-Bench.

3.1 Continuous Detection of Language-Defined Events

Before observing a video V_v of duration T_v, the model receives language-defined event candidates

$$\mathcal{Q}_v = \{q_i\}_{i=1}^{N_v}. \tag{1}$$

Each candidate has zero or more annotated temporal intervals,

$$\mathcal{G}_{vi} = \{[s_{vim}, e_{vim}]\}_{m=1}^{M_{vi}}, \qquad y_{vi} = \mathbf{1}[M_{vi} > 0]. \tag{2}$$

A candidate is positive if $M_{vi} > 0$ and negative otherwise; repeated occurrences are treated as distinct instances. During evaluation, only candidate descriptions are visible, while occurrence labels, temporal boundaries, difficulty labels, and verification metadata are hidden.

Causal monitoring. At every integer second $d_k = k$, $k = 1, \ldots, \lfloor T_v \rfloor$, the detector uses only observations available so far and its previous predicted state:

$$(\mathcal{A}_{v,k}, z_{v,k}) = f_\theta\big(\mathcal{O}_v(d_k), \mathcal{Q}_v, z_{v,k-1}\big), \qquad \mathcal{O}_v(d_k) \subseteq V_{v,\le d_k}. \tag{3}$$

$\mathcal{A}_{v,k}$ contains predicted `START`/`END` transitions or `NONE`. A `START` reports an onset, `END` updates the predicted state for later occurrences, and `NONE` leaves it unchanged. No ground-truth state is used during inference.

Per-second observation protocol. For shared-window VLMs, the target frame times at second k are

$$\mathcal{U}_k = \{\max(0, k-4), \ldots, k\}. \tag{4}$$

Each target uses the latest source frame no later than that time, yielding at most five recent frames at 1 fps; the first decision observes $\{0, 1\}$ seconds. All candidates share one model call, and no additional decision is made for a final fractional second. Accepted `START` and `END` transitions activate and deactivate candidates, respectively; end-of-video closure only organizes predicted intervals. Native streaming systems follow the same causal task with model-specific interfaces detailed in Appendix D.

3.2 Evaluation Metrics

CEDVS-Bench evaluates three capabilities of continuous monitoring: accurate onset localization, timely reporting, and reliable rejection of absent events. We further distinguish when a decision is made from when its response becomes available.

Onset localization. Each accepted `START` report $r \in \mathcal{R}_{vi}$ records predicted onset $\hat{s}_r$, decision time d_r, and simulated completion time c_r. For tolerance $\delta \in \{1, 2, 3\}$ seconds, **OED (onset localization)** considers a prediction correct if

$$|\hat{s}_r - s| \leq \delta, \tag{5}$$

where s is the ground-truth onset. OED evaluates localization regardless of when the report is produced.

Decision timeliness. A correctly localized onset may still be reported too late. **DT (decision timeliness)** therefore requires

$$0 \leq d_r - s \leq \delta. \tag{6}$$

Matching first maximizes the number of valid matches and then minimizes total delay. For OED and DT, predictions are matched one-to-one to ground-truth onsets, and we report micro-averaged precision, recall, and F_1. Reports on absent candidates and unmatched duplicate `START`s count as false positives; detailed matching rules are provided in Appendix C.OED and DT use the evaluator's implemented matching rules described in Appendix C; because their matching objectives differ, DT is not numerically constrained to be lower than OED.

Response timeliness. A timely decision may still become available too late because of processing cost. Given measured service time S_k, we simulate serial FIFO execution:

$$B_k = \max(d_k, C_{k-1}), \qquad C_k = B_k + S_k, \qquad C_0 = 0. \tag{7}$$

For shared-window VLMs, reports at step k have completion time $c_r = C_k$. **SRT (simulated response timeliness)** uses the DT matching objective with

$$d_r \geq s, \qquad 0 \leq c_r - s \leq \delta. \tag{8}$$

Thus, a report must be decided after event onset and become available within the tolerance, implying $F_{1,\mathrm{SRT}} \leq F_{1,\mathrm{DT}}$.

Service time includes input preparation, inference, output parsing, and state updates. We additionally report mean/P95 service time and

$$\mathrm{RTF} = \frac{\sum_k S_k}{\lfloor T_v \rfloor \times 1\mathrm{s}}. \tag{9}$$

These quantities support cost-based replay rather than measured live end-to-end latency.

Negative-event rejection. For absent candidates $\mathcal{N}_\ell$ at difficulty level ℓ, we measure candidate-level **false-alarm rate (FAR)**:

$$\mathrm{FAR}_\ell = \frac{\sum_{(v,i)\in\mathcal{N}_\ell} \mathbf{1}[|\mathcal{R}_{vi}| > 0]}{|\mathcal{N}_\ell|}. \tag{10}$$

A candidate fails rejection if it produces any false `START` during the video. FAR is reported with positive-event recall to avoid rewarding trivial silence.

Table 2: Statistics of the three CEDVS-Bench subsets. Candidates are unique event descriptions within each video; positive instances correspond to retained temporal intervals. Counts use retained evaluation annotations, and mean duration is computed from media metadata.

Subset	Videos	Pos. cand.	Pos. inst.	Neg. cand.	Mean sec.
CEDVS-Short	1000	2432	2482	6887	29.4
CEDVS-Medium	1487	7316	7599	45496	149.6
CEDVS-Long	353	5765	5902	13775	888.2

Table 3: Human assessment of audited samples retained after automatic filtering. Positive validity uses all retained positive records; negative validity excludes unresolved judgments. The pooled row aggregates audited records without subset weighting.

Subset	Videos	Retained pos. / neg.	Positive validity (%)	Negative validity (%)	Unresolved pos. / neg.
CEDVS-Short	40	102 / 280	83.33	97.13	1 / 1
CEDVS-Medium	40	184 / 1,040	96.74	97.87	0 / 7
CEDVS-Long	30	220 / 518	94.55	97.68	0 / 0
Pooled	**110**	**506 / 1,838**	**93.08**	**97.70**	**1 / 8**

Additional diagnostics. We additionally report conditional onset error and DT/SRT delays; detailed definitions are provided in Appendix C.

3.3 Dataset Construction

CEDVS-Bench contains three subsets—**CEDVS-Short**, **CEDVS-Medium**, and **CEDVS-Long**—covering substantially different video durations while sharing the same task definition, annotation format, and evaluation protocol. Rather than designing separate tasks for different video sources, we apply a common construction procedure that converts videos into continuous monitoring examples composed of language-defined positive and absent event candidates.

Unified construction pipeline. As shown in Figure 2, all three subsets follow the same construction stages: video selection, visual description, event extraction and consolidation, temporal annotation, absent-event construction, automatic verification, and human quality inspection.

For each video, the construction process produces a set of language-defined event candidates. Positive candidates describe events observed in the video and are associated with one or more temporal intervals. Distinct event occurrences are retained as separate instances. Negative candidates describe plausible events that are absent throughout the video, enabling the benchmark to evaluate whether a model can consistently reject semantically related but unsupported events. Automatic verification is used to filter inconsistent candidates and annotations. Human inspection independently evaluates positive-event validity, negative-event validity, and temporal agreement. Detailed prompts, model versions, filtering rules, adjudication procedures, and quality-control protocols are provided in Appendices A and B.

Dataset composition and quality. The resulting benchmark is organized into three subsets according to their characteristic video duration. Table 2 summarizes their scale. CEDVS-Short contains shorter videos with a mean duration of 29.4 seconds, CEDVS-Medium covers videos averaging 149.6 seconds, and CEDVS-Long extends continuous monitoring to videos averaging 888.2 seconds. Across these subsets, the same benchmark protocol is preserved while the observation horizon and number of monitored events vary substantially. To assess annotation quality after automatic filtering, we further conduct human inspection on sampled videos from all three subsets. Table 3 reports the retained positive and negative records together with their validity rates. Across the audited samples, positive validity reaches 93.08% and negative validity reaches 97.70%, with only a small number of unresolved judgments. These results indicate that the automatic construction and filtering pipeline retains high-quality event annotations while supporting large-scale benchmark construction.

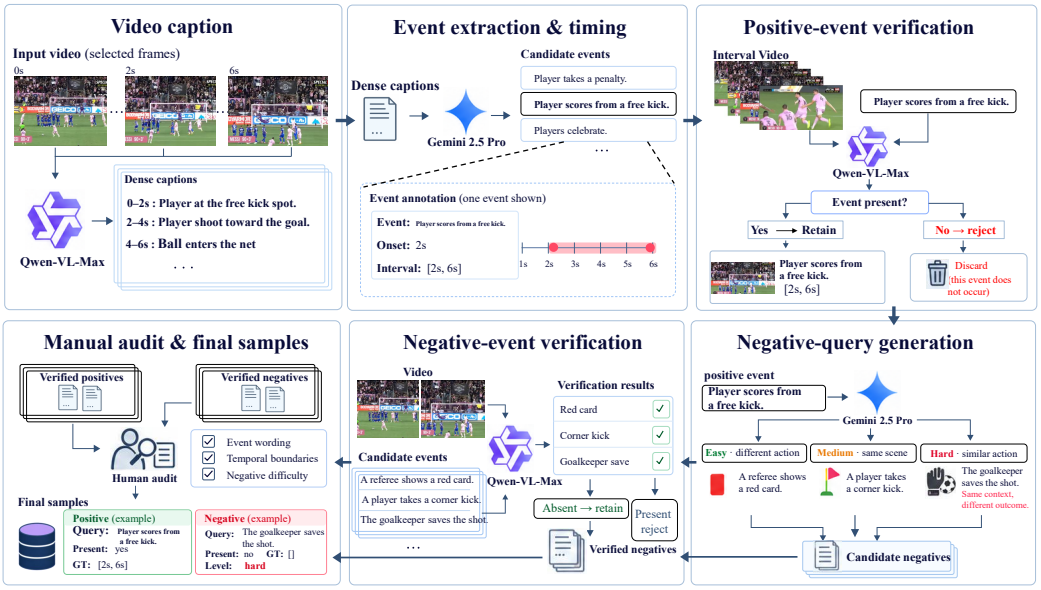

Figure 2: Unified construction pipeline of CEDVS-Bench. The same pipeline is applied across different video durations, including visual description, event extraction and consolidation, temporal annotation, absent-event construction, automatic verification, and human quality inspection.

The three subsets therefore provide complementary temporal settings within a single benchmark: CEDVS-Short focuses on relatively compact event streams, CEDVS-Medium extends monitoring over longer contexts, and CEDVS-Long evaluates sustained detection and rejection over substantially longer observation periods. Importantly, the task itself remains unchanged across subsets: event candidates are specified before observation, decisions are causal, and both event triggering and continued silence are evaluated throughout the stream. Further statistics on event distributions, domain coverage, negative-event difficulty, and detailed quality-assessment procedures are provided in Appendix B.

4 Experiments

4.1 Experimental Setup

We evaluate eight general VLMs with causal recent windows (Bai et al., 2025; Wang et al., 2025a; Yao et al., 2024; An et al., 2025; Zhang et al., 2024b; 2025; 2024a; Shen et al., 2024): one-second decisions, at most five recent frames, and no benchmark-specific training. Dispider, VideoLLM-Online, and M4 form a separate native streaming family (Qian et al., 2025; Chen et al., 2024; Wang et al., 2025b); their adapters require separate timing and coverage checks (Appendix D). Appendix G varies visual context and candidate grouping.

Scores are micro-averaged over all videos in the corresponding CEDVS subset and expressed as percentages; – denotes settings that were not evaluated. Appendices D–E provide detailed model configurations, streaming interfaces, complete metrics, and efficiency results. Offline conditions align videos, candidates, and true intervals. Response scores replay recorded preparation and inference costs; heterogeneous hardware and load preclude controlled architecture-speed comparisons.

4.2 Experiment 1: Accurate and Timely Continuous Detection

Accurate localization does not ensure timely availability. Qwen leads the shared-window block in onset F1 on all three sources (Table 4). Its CEDVS-Short decision F1 is 39.03, versus 14.62 under response scoring; every shared-window model has CEDVS-Long response F1 below 0.1. Localiza-

tion and service-aware availability therefore expose distinct limitations. Native and shared-window results are not directly comparable because candidate presentation, observation cadence, state maintenance, and timing instrumentation differ. M4 reaches an onset F1 of 13.89 on CEDVS-Short, while Dispider and VideoLLM-Online remain below 3 across sources. These results reflect the supplied implementations and should be interpreted together with the interface differences documented in Appendix D.

Table 4: Main F1@3s (%). Onset: localization; Decision: timely detection; Response: simulated timely availability. All reported shared-window results are evaluated on the full corresponding subset. Native streaming interfaces and unavailable settings are detailed in Appendix D. –: not evaluated due to model-specific interface or resource constraints; see Appendix D.

	CEDVS-Short			CEDVS-Medium			CEDVS-Long		
Model	Onset	Decision	Response	Onset	Decision	Response	Onset	Decision	Response
General VLMs with causal recent windows									
Qwen3-VL-8B	41.75	39.03	14.62	19.88	21.05	9.75	7.78	7.00	0.06
InternVL3.5-8B	23.19	21.42	5.38	7.96	7.74	0.67	3.92	3.18	0.04
MiniCPM-V-2.6	24.21	19.40	3.91	6.76	5.13	0.04	2.51	1.82	0.03
LLaVA-OneVision-1.5-8B	23.96	22.09	1.65	10.07	9.94	0.03	3.72	3.35	0.00
LLaVA-NeXT-Video-7B	12.14	10.86	1.82	4.22	4.00	< 0.01	2.59	2.50	0.00
VideoLLaMA3-7B	22.09	18.88	2.81	5.29	4.41	0.00	3.10	2.51	0.00
LongVA-7B	15.73	12.49	1.31	7.07	5.89	< 0.01	4.55	4.02	0.00
LongVU-7B	32.75	27.73	4.26	10.24	8.00	0.01	3.33	2.70	0.00
Native streaming video models									
Dispider	2.87	1.93	1.68	1.67	0.77	0.54	0.46	0.21	0.18
VideoLLM-Online	1.40	1.25	1.18	1.08	1.08	0.92	0.19	0.19	0.12
M4	13.89	2.01	1.72	2.58	0.58	0.50	–	–	–

4.3 Experiment 2: Recognition, Localization, and Streaming Triggers

Recognition rankings need not transfer to localization. Offline inference sees a sampled full video and returns intervals or NONE; presence and onset F1 use the same output. On CEDVS-Medium, InternVL exceeds Qwen in presence F1 (62.76 versus 56.63) but trails in offline onset F1 (16.70 versus 37.18; Table 5). Qwen's CEDVS-Short presence, offline onset, and online onset scores likewise separate these abilities.

Explicit frame timestamps reduce onset F1 in all eight tested model–source pairs, under these prompts and sampling policies. MiniCPM scores higher online on both subsets, while VideoLLaMA3 does so on CEDVS-Short. Horizon, sampling, prompts, and call frequency change together, preventing attribution to future-frame access alone. Malformed completed answers remain no-detection outputs.

4.4 Experiment 3: Rejecting Similar but Absent Events

Detection and rejection must be assessed jointly. Qwen's CEDVS-Medium FAR rises from 13.38% for Easy negatives to 45.04% for Hard negatives, despite leading shared-window onset F1 (Table 6). Similarity therefore challenges rejection. A small Hard–Easy gap can instead reflect saturation: LLaVA-NeXT triggers on essentially every negative, and VideoLLaMA3 exceeds 98% FAR at each difficulty. Low FAR also requires recall context: CEDVS-Long LongVA achieves 18.92% FAR but only 4.26% recall (Appendix E). Aggregate scores establish false triggering; identifying entity, relation, or state errors requires verified cases (Appendix H).

4.5 Experiment 4: Reliability from Short to Long Videos

Longer exposure accumulates rejection failures. Both models' onset recall decreases across CEDVS-Short, CEDVS-Medium, and CEDVS-Long, but duration is confounded with domain, can-

Table 5: Offline and online F1 (%), with 3s temporal tolerance. The same four models and all three conditions are evaluated on matched videos. Presence and offline onset scores come from the same output.

Model	Offline presence	Offline onset	Offline onset with timestamps	Online onset	Online decision
CEDVS-Short					
Qwen3-VL-8B	78.48	59.11	57.50	41.75	39.03
InternVL3.5-8B	65.08	41.89	40.43	23.19	21.42
MiniCPM-V-2.6	46.43	20.86	16.34	24.21	19.40
VideoLLaMA3-7B	19.49	14.57	14.08	22.09	18.88
CEDVS-Medium					
Qwen3-VL-8B	56.63	37.18	36.43	19.88	21.05
InternVL3.5-8B	62.76	16.70	13.80	7.96	7.74
MiniCPM-V-2.6	15.46	4.95	4.37	6.76	5.13
VideoLLaMA3-7B	18.34	6.62	3.64	5.29	4.41

Table 6: CEDVS-Medium positive onset Recall@3s and negative FAR (%). H−E is the Hard minus Easy gap in percentage points. Coverage follows Table 4; low gaps can reflect uniformly high false alarms.

Model	Recall ↑	Easy FAR ↓	Medium FAR ↓	Hard FAR ↓	H−E
Qwen3-VL-8B	38.27	13.38	31.62	45.04	+31.66
InternVL3.5-8B	27.40	67.91	86.07	92.44	+24.53
MiniCPM-V-2.6	29.49	76.46	81.68	85.47	+9.01
LLaVA-OneVision-1.5-8B	31.70	53.97	73.55	82.74	+28.77
LLaVA-NeXT-Video-7B	18.00	100.00	100.00	99.99	-0.01
VideoLLaMA3-7B	22.02	98.23	98.49	98.59	+0.36
LongVA-7B	24.21	67.04	67.28	71.45	+4.42
LongVU-7B	21.88	40.95	42.48	50.38	+9.42

didate load, and event granularity. We therefore examine the same CEDVS-Long videos and candidates at 30s, 150s, and full-video horizons. Reports are truncated by decision time; future true onsets are censored, not relabeled absent. Alongside FAR, we measure negative START frequency per query-minute:

$$\mathrm{FA}_{\text{query-min}}(h) = \frac{\sum_v A_v(h)}{\sum_v n_v^- h_v/60}, \tag{11}$$

where $A_v(h)$ includes repeated negative STARTs and h_v is the integer-second horizon. From 30s to the full video, Qwen's FAR rises from 14.65% to 68.16%, while alarm frequency falls from 0.314 to 0.051; InternVL shows the same directions (Table 7). More candidates are eventually misidentified despite fewer false STARTs per unit exposure. Decision F1 also declines, but eligible true onsets grow from 588 to 5,696, preventing a fixed-event accuracy interpretation.

Table 7: Fixed-video exposure on the CEDVS-Long videos for both models. Entries are Qwen / InternVL. F1 and FAR are percentages; every video is longer than 150s.

Horizon	Visible onsets	Decision F1@3s	Negative FAR	FA/query-min
First 30s	588	14.04 / 7.50	14.65 / 29.33	0.314 / 0.587
First 150s	1445	9.62 / 4.19	34.73 / 68.00	0.153 / 0.272
Full video	5696	7.00 / 3.18	68.16 / 85.92	0.051 / 0.056

4.6 Experiment 5: Detection across Event Durations

Each true occurrence has duration $D_{vim} = e_{vim} - s_{vim} > 0$, with fixed bins

$$\text{Short} : 0 < D \leq 8\text{s}; \quad \text{Medium} : 8 < D \leq 30\text{s}; \quad \text{Long} : D > 30\text{s}. \tag{12}$$

We match globally before assigning true instances to bins, retaining missed occurrences in recall denominators. Repeated occurrences are not summed. Unmatched predictions have no true-instance duration, so per-bin F1 would require an arbitrary false-positive assignment.

Long events are not uniformly easier. Qwen's CEDVS-Short onset recall rises from 52.26% for Short to 90.91% for Long events; CEDVS-Medium shows no such increase, and CEDVS-Long Long recall is 14.95% (Table 8). CEDVS-Short has only 66 Long instances, occupying a median 98.0% of their videos, versus 7.8% for CEDVS-Long Long instances. Absolute duration therefore represents different observation conditions across sources. Within-source, within-duration median splits further reveal positive occupancy–recall associations throughout CEDVS-Short but mixed signs elsewhere (Figure 3). These are descriptive associations, not isolated occupancy effects. Near-complete CEDVS-Long coverage coexists with low recall and large onset error: a START for a candidate does not establish accurate localization of its occurrences. Appendix F provides the detailed results.

Table 8: True-instance duration analysis on identical Qwen/InternVL video sets. Recall and coverage are percentages; conditional onset error is in seconds.

Duration	Instances	Onset recall@3s	Decision recall@3s	Onset error (s)	Coverage
CEDVS-Short					
Short	1263	52.3 / 38.5	49.7 / 36.0	5.3 / 9.1	83.1 / 98.9
Medium	1153	65.6 / 57.2	60.6 / 52.6	3.2 / 5.1	90.5 / 99.8
Long	66	90.9 / 92.4	80.3 / 81.8	0.8 / 0.8	97.0 / 100.0
CEDVS-Medium					
Short	1685	41.5 / 23.4	46.2 / 24.3	24.2 / 44.4	91.8 / 99.5
Medium	3853	37.3 / 26.2	38.6 / 25.2	25.8 / 41.0	91.9 / 99.6
Long	2061	37.3 / 32.9	39.4 / 31.2	21.1 / 31.2	92.9 / 99.7
CEDVS-Long					
Short	917	17.1 / 8.2	18.4 / 6.7	255.2 / 336.8	99.5 / 99.1
Medium	2344	17.4 / 8.3	16.2 / 6.5	267.9 / 373.7	99.0 / 99.6
Long	2435	14.9 / 7.4	11.8 / 6.0	266.4 / 359.9	99.4 / 99.7

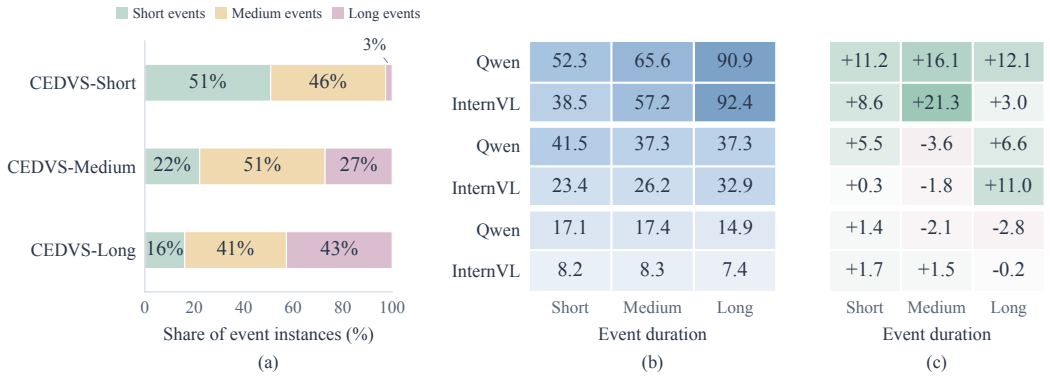

Figure 3: Instance-level analysis on the matched cohorts of Table 8. (a) Duration composition; (b) onset recall for both models on all sources; (c) recall difference between higher- and lower-occupancy instances, in percentage points. Occupancy is event/video duration; each source and duration bin uses its own annotation median, with ties assigned lower. No causal effect is implied.

5 Conclusion And Future Work

In this paper, we introduce CEDVS-Bench to evaluate continuous monitoring of language-defined events under causal video observation. By jointly assessing onset localization, reporting timeliness, and sustained rejection, the benchmark reveals gaps that retrospective recognition scores alone cannot capture. Across the evaluated models, recognizing relevant content does not consistently translate into accurate, timely event reports or reliable rejection of plausible alternatives. These findings motivate future work on fine-grained event verification, temporal state tracking, and adaptive computation that balances detection quality with response deadlines. Expanding boundary-focused human validation and evaluating dynamic instructions, audiovisual evidence, and live execution would further strengthen the benchmark's scope and practical relevance. Together, these directions support progress toward video assistants that respond when warranted and remain silent otherwise.

AI USE STATEMENT

Generative AI tools assisted language editing, manuscript organization, LaTeX, and the overview figure. The authors checked the experimental code, data, and manuscript for correctness and consistency and take responsibility for the final work.

ETHICS STATEMENT

CEDVS-Bench uses existing publicly available source videos and does not involve collecting new video recordings of individuals. Public availability does not eliminate potential privacy concerns. Continuous video monitoring may also have application-dependent consequences when systems produce false alarms or miss relevant events. The benchmark is intended for research on these capabilities and limitations; its results do not establish suitability for safety-critical deployment. We encourage responsible use and adherence to the licenses and usage terms of the original video sources. Any redistribution of source videos will be subject to the applicable permissions.

REPRODUCIBILITY STATEMENT

We will publicly release all CEDVS-Bench dataset files and annotations, together with the complete codebase for dataset construction, automatic verification, benchmark evaluation, and reproduction of the reported experiments, including model configurations and experiment settings. The paper and appendices provide the task definition, evaluation protocol, and main construction details, while the released codebase will contain the complete implementation.

REFERENCES

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Wentao Bao, Qi Yu, and Yu Kong. Opental: Towards open set temporal action localization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2969–2979. IEEE, 2022.

Dibyadip Chatterjee, Zhanzhong Pang, Fadime Sener, Yale Song, and Angela Yao. Don't pause! every prediction matters in a streaming video. *arXiv preprint arXiv:2604.24317*, 2026.

Guo Chen, Yicheng Liu, Yifei Huang, Baoqi Pei, Jilan Xu, Yuping He, Tong Lu, Yali Wang, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. In *International Conference on Learning Representations*, volume 2025, pp. 45647–45682, 2025.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18407–18418. IEEE, 2024.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5277–5285. IEEE, 2017.

Akshita Gupta, Aditya Arora, Sanath Narayan, Salman Khan, Fahad Shahbaz Khan, and Graham W Taylor. Open-vocabulary temporal action localization using multimodal guidance. *arXiv preprint arXiv:2406.15556*, 2024.

Zhenpeng Huang, Xinhao Li, Jiaqi Li, Jing Wang, Xiangyu Zeng, Cheng Liang, Tao Wu, Xi Chen, Liang Li, and Limin Wang. Online video understanding: Ovbench and videochat-online. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3328–3338. IEEE, 2025.

Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021.

Junming Lin, Zheng Fang, Chi Chen, Haoxuan Cheng, Zihao Wan, Fuwen Luo, Ziyue Wang, Peng Li, Yang Liu, and Maosong Sun. Streamingbench: Assessing the gap for mllms to achieve streaming video understanding. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12147–12151. IEEE, 2026.

Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Ying Shan, and Chang W Chen. Et bench: Towards open-ended event-level video-language understanding. *Advances in Neural Information Processing Systems*, 37:32076–32110, 2024.

Junbo Niu, Yifei Li, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, et al. Ovo-bench: How far is your video-llms from real-world online video understanding? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18902–18913. IEEE, 2025.

Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24045–24055. IEEE, 2025.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Yuxuan Wang, Yueqian Wang, Bo Chen, Tong Wu, Dongyan Zhao, and Zilong Zheng. Omni-immi: A comprehensive multi-modal interaction benchmark in streaming video contexts. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18925–18935. IEEE, 2025b.

Zhenyu Yang, Yuhang Hu, Zemin Du, Dizhan Xue, Shengsheng Qian, Jiahong Wu, Fan Yang, Weiming Dong, and Changsheng Xu. Svbench: A benchmark with temporal multi-turn dialogues for streaming video understanding. In *International Conference on Learning Representations*, volume 2025, pp. 42522–42556, 2025.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024a.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024b. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

Yusong Zhao, Yuejin Xie, Youliang Yuan, Junjie Hu, Jitian Guo, Yujiu Yang, and Pinjia He. Pasbench-video: A streaming video benchmark for proactive safety warning. *arXiv preprint arXiv:2606.02443*, 2026.

A Dataset Construction

CEDVS-Bench applies a shared construction pipeline to all three subsets. The pipeline converts source videos into persistent monitoring tasks, each consisting of natural-language event candidates and their video-level occurrence labels. A positive candidate is associated with one or more temporal intervals; an absent candidate specifies an event intended to be absent throughout the recording. The construction process has access to the complete video to establish these labels. This access is confined to annotation: evaluated models receive only causal visual observations and the candidate descriptions, as specified in Section 3.1. Figure 2 summarizes the stages; the following sections detail how the descriptions, temporal intervals, and rejection targets are obtained.

A.1 Video Selection and Visual Description

The videos of CEDVS-Bench are sourced from Charades (Gao et al., 2017), QVHighlights (Lei et al., 2021) and CGBench (Chen et al., 2025). We construct the benchmark from available videos associated with the selected source splits, resolving source identifiers to local media and checking that the videos can be decoded and have valid duration metadata. Repeated references to the same source video are consolidated before processing. The long-video source pool is restricted to recordings shorter than twenty minutes. Source identifiers and category metadata support provenance and coverage analysis; the event-generation stages operate on visual descriptions rather than treating the source question–answer pairs as event annotations. Table 2 reports the retained benchmark, rather than the sizes of the original sources.

Each video is divided into consecutive clips on a two-second temporal grid, with the final clip bounded by the actual video duration. We sample two temporally ordered frames within each clip and use Qwen-VL-Max to produce one concise description of its visible content. A small context of preceding captions maintains consistent references to participants and objects; the caption must describe the current clip rather than repeat an action supported only by earlier clips. The prompt emphasizes observable actions, interactions, and state changes, and discourages inferences about hidden intentions, unseen causes, or outcomes not supported by the frames. A clip without a meaningful action may receive a factual description of its visible state; the model is not instructed to invent an event for every clip.

The resulting record associates each caption with its clip identifier and temporal extent. These timestamped descriptions form a chronological representation of the whole video. Unavailable captions remain explicitly marked as missing evidence: they are not interpreted as evidence that an event stopped or that a candidate was absent. The dense-caption timeline is used only during construction and is not supplied to evaluated models.

A.2 Positive Event Extraction and Consolidation

Gemini-2.5-Pro converts the ordered caption timeline into structured event proposals. Each proposal contains a concise event sentence, a longer supporting description, proposed start and end times, and the identifiers of the clips that support it. The event sentence is written as a natural subject–verb–object or subject–activity expression, so that it can be understood as a monitoring condition without references such as "this clip" or "the preceding moment." The supporting description and evidence identifiers remain construction metadata, rather than additional information available to the detector.

The extractor groups successive observations into temporally coherent activities. For example, retrieving ingredients, cutting vegetables, and combining them may support one event, "A woman prepares food," rather than several candidates differing only in individual preparation steps. A change in the primary activity, interaction, or semantic episode supports a new event; a camera change or a different manipulated object does not, by itself, require a new boundary. The reference extraction configurations impose a six-second minimum on automatically proposed macro-events to discourage isolated clip-level fragments. This constraint concerns event proposal generation; the event-duration bins used for evaluation are defined separately in the main text.

A second language-model pass reviews adjacent proposals together with the original timeline. It merges proposals when they represent phases of the same ongoing activity, retaining their combined supporting evidence and temporal extent. The consolidation prompt forbids inventing unsupported

activities or splitting existing proposals into finer actions. Sharing a participant or location alone is insufficient for merging: preparing a meal and eating it can remain different events. Temporally separate occurrences remain distinct intervals, even when their final event descriptions identify the same monitoring condition. This semantic consolidation precedes the evaluator's candidate-identity and interval-deduplication rules described in Appendix B.

A.3 Temporal Annotation and Positive Verification

Origin of temporal boundaries. Temporal intervals are generated from the timestamped caption evidence during event extraction and consolidation; they are not presented as independently hand-annotated boundaries or as a direct conversion of the source datasets' temporal labels. The implementations normalize proposed endpoints to the clip grid or reconstruct the interval from the first and last supporting clips. For an occurrence with supporting clip set E and clip boundaries (a_j, b_j), the evidence-based reconstruction is

$$s = \min_{j \in E} a_j, \qquad e = \max_{j \in E} b_j.$$

Endpoints are constrained to the video duration, and an interval must satisfy $0 \leq s < e \leq T_v$. Malformed proposals, unusable evidence references in the evidence-based reconstruction, and intervals failing the generation-time duration constraint are excluded. Missing captions between supporting observations do not automatically split a continuing event. The procedure does not require every second of a video to belong to an event.

What the onset represents. The onset is the beginning of the annotated activity interval at the construction resolution. For a consolidated activity, this can include its visually supported preparation and execution phases; it is not necessarily the instant at which a final outcome becomes observable. The annotation timestamp is therefore distinct from the detector's decision time and its response completion time. Clip-grid normalization alone does not establish frame-accurate onset annotation, and we do not claim an additional universal boundary-refinement pass beyond the documented construction and sampled boundary review.

Visual support checking. Qwen-VL-Max checks whether the proposed event is supported by frames from its annotated interval. Frames are selected preferentially from the evidence clips, with uniform sampling within the interval as a fallback and at most twelve frames per event. The verifier is instructed to identify the core action or interaction; the presence of a related person, object, or scene is insufficient. It records a Boolean judgment and a reason, while unsuccessful calls remain distinguishable from semantic rejection. This check assesses visual support for the event, not independently measured boundary accuracy. The retained export and its inclusion rules are documented in Appendix B.

A.4 Absent-Event Construction

An absent candidate is an affirmative event description whose occurrence label is negative for the *entire* video. Being absent from the interval of its associated positive event is insufficient: a candidate that occurs elsewhere in the recording is an invalid negative. We construct three categories that vary their proximity to the observed content. The categories are defined by their construction rules, not by how frequently a particular evaluated model triggers on them.

Easy: cross-video event sampling. Easy candidates are sampled from an event pool drawn from other videos, excluding proposals explicitly rejected by positive-event verification. The sampler prioritizes descriptions with little content-word overlap with the target video's events; if this pool is insufficient, it can fall back to other cross-video events. Thus, Easy negatives are natural event descriptions rather than syntactic negations or random word substitutions. Their origin in a different video is a sampling heuristic, not a logical guarantee of absence from the target video.

Medium: context-preserving activity changes. Gemini-2.5-Pro generates Medium candidates conditioned on the full caption timeline and the video's event list. The candidate should be plausible in the same scene or broad activity context while changing a participant, core activity, action, or

interaction. For example, "A woman washes dishes" may be a Medium alternative to "A woman prepares food," but only if dishwashing is unsupported throughout the target recording. The generator is required to consult the whole timeline, rather than only the positive event used to initiate the proposal.

Hard: minimal changes to the event condition. Hard candidates preserve the participant, scene, and much of the event's semantic structure, while changing one or a few critical facts. The generation prompt specifies transformations such as changing a closely related action, replacing the manipulated object, reversing a participant relation, or changing an outcome. Illustrative transformations include "A woman cuts vegetables" to "A woman cuts fruit," and "A man gives an object to a woman" to "A woman gives an object to a man." These are examples of the construction rule, not automatically valid negatives: each proposed alternative must still satisfy the full-video absence criterion. This design tests whether a detector grounds the particular event condition rather than responding to a shared person, object, or setting.

Constraints shared by generated alternatives. Medium and Hard candidates must match the concise affirmative style of positive descriptions; explicit negations such as "does not" and "without" are disallowed. The prompt excludes paraphrases and broader statements entailed by a positive event. For example, "A person touches a cup" is not a valid negative solely because the positive is phrased as "A person picks up a cup." It also requests distinct alternatives within each category and a short rationale explaining why each alternative is unsupported by the supplied timeline. Generation can be batched over positive events, but the context includes the whole timeline and the video's event list to reduce collisions with events elsewhere in the recording. Rationales and difficulty labels are retained as metadata and are not exposed to evaluated models.

A.5 Video-Wide Negative Verification and Consistency Checks

Searching for counterevidence. Medium and Hard proposals undergo a separate visual check with Qwen-VL-Max. For each associated positive event, the verifier receives up to sixteen uniformly spaced frames covering the video and four additional samples within the positive interval; nearby duplicate timestamps are removed. This combines broad temporal coverage with additional evidence near the event from which a semantically similar alternative was generated. Candidates associated with that event share the sampled visual evidence and are checked together. The prompt asks the verifier to search for evidence that a purportedly absent event *actually occurs*, rather than merely endorse the generator's rationale.

The verifier returns `yes`, `no`, or `unsure` for occurrence, together with a reason and, when occurrence is supported, an evidence timestamp. These outcomes are recorded as `removed`, `kept`, and `unsure`, respectively. A missing API result remains a separate unresolved state. Shared scene context is not sufficient counterevidence: removal requires support for the candidate's core action or interaction. Easy candidates are not subjected to this additional visual-verification call in the implemented pipeline; their automatic `kept` status must not be interpreted as the same visual check used for Medium and Hard candidates.

Structured validation and export. The construction code checks required fields, parses temporal values, constrains intervals to valid video times, and orders event records chronologically. Generation-time duplicate checks and prompt constraints reduce repeated alternatives, but they do not constitute a guarantee that all semantic paraphrases have been eliminated. Verification states are preserved for traceability rather than silently converting ambiguous judgments into confirmed labels. The exact candidate inclusion, conflict precedence, and deduplication rules of the retained evaluation exports are specified once in Appendix B; they should be distinguished from the intermediate generation and verification records described here.

Scope of the absence claim. Full-video absence is the target annotation criterion, whereas the automated procedure supplies finite evidence toward that criterion. The generator considers descriptions spanning the recording, and the visual verifier samples across its full duration, but neither constitutes exhaustive frame-by-frame verification. A brief or poorly described occurrence can be missed, and cross-video Easy candidates can also coincide with an event in the target video. Ac-

cordingly, automatic verification is complemented by sampled human inspection rather than used to claim that every retained negative is certainly absent.

A.6 Human Quality Inspection and Final Evaluation Interface

Human inspection evaluates whether a positive description is faithful to an observed event, whether a negative candidate is absent throughout the video, and whether annotated temporal boundaries are acceptable or require correction. It is a sampled quality audit, not exhaustive manual annotation of the benchmark. Table 3 reports the actual validity results; Appendix B.4 defines their denominators and unresolved judgments, and Table B.1 reports the boundary review. These existing results are not duplicated here. In particular, accepted boundaries remain unchanged in that review: the reported differences measure an accept-or-correct procedure, not agreement between independently produced temporal annotations. Difficulty-stratified absence validity likewise does not establish human agreement with the assigned difficulty categories.

The final monitoring interface provides the candidate descriptions without revealing their positive/negative labels, temporal intervals, difficulty categories, caption evidence, or verification rationales. Under the joint-candidate protocol, all descriptions are supplied before observation begins; any model-specific grouping changes how candidates are presented, not which annotation metadata are available. Positive intervals determine event-level evaluation targets, while absent candidates support whole-video rejection analysis. The shared construction stages thereby yield a common monitoring format across sources, with dataset statistics and empirical quality evidence reported separately in Appendix B.

A.7 Prompts for Event Construction and Verification

We provide the negative-generation, positive-verification, and negative-verification instructions used by the construction pipeline below. The templates are typeset for readability: runtime inputs appear as labeled placeholders, and each output schema illustrates a single record rather than repeating identical entries. Negative-generation and negative-verification instructions follow the implementation templates; the Chinese positive-verification user instruction is presented as an English translation. These prompts accompany the visual and textual inputs described above. They are construction instructions and are not supplied to the models evaluated on CEDVS-Bench.

Negative-event generation. Gemini-2.5-Pro receives the complete caption timeline and the video's positive-event list. The requested number of alternatives per level is a runtime parameter. Easy negatives are sampled from other videos, so this prompt generates only Medium and Hard alternatives. The task instructions and output schema are shown separately to keep both readable.

Prompt 1 Negative-event generation: task and constraints

You are creating NEGATIVE samples (distractor event descriptions) for a video understanding dataset. A negative sample is an event that did NOT happen anywhere in the video.

Video context. Here is the full caption timeline of a `{duration}`-second video (one caption per 2s clip): `{timeline}`. Here are the real events extracted from this video: `{events_list}`.

Task. For EACH event listed above, write `{n_per_level}` "medium" negative samples and `{n_per_level}` "hard" negative samples:

Medium: keeps the SAME scene/environment as the event, but changes the person, the action, or the event itself to something different.

Hard: keeps the SAME scene AND the same person(s), and is semantically very similar to the real event, but describes something that did NOT actually happen (e.g., a closely related but different action, the reverse action, or a different object).

Strict rules for every negative sample.

1. It must NOT describe anything that actually happens ANYWHERE in the video. Check the full caption timeline before writing; if the candidate matches any part of the video, pick a different one.
2. It must NOT be a paraphrase or a more general/more specific version of the real event. If the real event is true, the negative must be clearly false.
3. Write an affirmative declarative sentence: no "does not," "fails to," or "without."
4. Match the style of the real events: a natural subject–verb–object phrase with articles and similar length. The falseness must be visually checkable from frames.
5. Samples within the same level for the same event must be clearly DIFFERENT from each other (different persons/actions/objects), NOT paraphrases of one another.
6. Give a short reason for each, explaining what in the video contradicts it or why it never happens.

Prompt 1 Negative-event generation: output schema

Return ONLY valid JSON, with no Markdown fences. Include one entry per event, using the supplied `event_index`, with exactly `{n_per_level}` items in each of the `medium` and `hard` arrays. The layout below shows one event and one item per array.

```
{
  "negatives": [
    {
      "event_index": 0,
      "medium": [
        {"text": "...", "reason": "..."}
      ],
      "hard": [
        {"text": "...", "reason": "..."}
      ]
    }
  ]
}
```

Positive-event verification. Qwen-VL-Max receives chronologically ordered frames sampled from the proposed positive interval and the candidate event name. The English system instruction and the translation of the Chinese user instruction are distinguished below. This check verifies visual support; it does not request independent temporal annotation or a new onset estimate.

Prompt 2 Positive-event verification

System instruction. You are a strict video event verification assistant. You will receive chronologically ordered frames sampled from one annotated time interval of a video, along with a proposed event name. Judge only from the visible evidence in the supplied frames.

User instruction *(translated from Chinese).* Determine whether the candidate event occurs in the supplied, chronologically ordered video frames.

Decision rules.

1. `verified=true`: clear, direct visual evidence supports the candidate's core action or interaction.
2. `verified=false`: the frames clearly contradict the event, or the available evidence is insufficient to support it.
3. The same person, object, or scene is insufficient to establish occurrence; the core action must be visible.
4. If the event name contains multiple actions, the core event must hold. Return false if a critical action is clearly missing.
5. Give a concise reason in Chinese, no more than 80 Chinese characters.
6. Return only one valid JSON object, with no Markdown, code fences, or additional text.

Output schema (one possible judgment).

```
{
  "verified": true,
  "reason": "<brief Chinese rationale>"
}
```

Negative-event verification. Qwen-VL-Max receives timestamped frames covering the video, additional samples within the associated positive interval, and indexed Medium/Hard candidates. The prompt searches for evidence that contradicts a proposed absence label.

Prompt 3 Negative-event verification

You are checking NEGATIVE samples for a video dataset. Each statement below is CLAIMED to describe something that does NOT happen in this video. Your job is to hunt for evidence that any of them actually DOES happen.

The frames are sampled from the whole video, each labeled with its timestamp. Frames are sparse, so:

- Say `happened="yes"` ONLY if you see actual visual evidence of the described subject performing the described action; cite the timestamp.
- Scene or person similarity alone is NOT evidence: the statements are intentionally written about the same scene/people. Judge the ACTION and OBJECT.
- Say `happened="unsure"` if some frames partially suggest the statement but you cannot confirm the core action.
- Say `happened="no"` if no frame supports it.

Statements to check: `{statements}`.

Output requirements. Return ONLY valid JSON, with no Markdown fences. Include one entry for every supplied `neg_index`. The `happened` field must be `"yes"`, `"no"`, or `"unsure"`; `evidence_time` is a timestamp in seconds or `null`. The layout below illustrates one entry.

```
{
  "results": [
    {
      "neg_index": 0,
      "happened": "no",
      "evidence_time": null,
      "reason": "<one concise sentence>"
    }
  ]
}
```

The occurrence judgments `yes`, `no`, and `unsure` are stored as `removed`, `kept`, and `unsure`, respectively. A `no` judgment describes the evidence in the sampled frames; it does not prove absence from every unsampled moment. Easy candidates bypass this additional verification call. Human inspection and the retained-export rules are reported in the existing data-quality appendix.

B Dataset Coverage, Statistics, and Human Inspection

B.1 Content Coverage and Annotation Structure

The three derivatives follow the shared construction pipeline in Section 3.3. Here we characterize the retained evaluation exports rather than the complete source benchmarks. As in the category and temporal analyses of CG-Bench and SVBench (Chen et al., 2025; Yang et al., 2025), content coverage and temporal structure are displayed separately. Figures B.1–B.3 are computed from annotations and media metadata; they contain no model-performance results and do not use the 100/20-video ablation subsets.

Candidate identity follows the evaluator: exact descriptions after trimming outer whitespace, with positive membership taking precedence over a conflicting absent label. Identical intervals are deduplicated. All retained positive candidates are included; for the CEDVS-Long export only negative entries whose normalized verification status is `kept` enter the evaluated candidate set. Recorded difficulty memberships may overlap. Missing difficulty labels remain *unspecified*; they are not reconstructed from the text.

Figure B.1 aligns the number of videos and positive event instances by content category. CEDVS-Short is described at the source level as indoor activities and CEDVS-Medium as recordings/vlogs; CEDVS-Long retains its 13 detailed source categories. These granularities are explicitly separated rather than treated as a shared fine-grained taxonomy. Each video contributes once per recorded domain, and its positive instances follow that membership.

B.2 Temporal Structure

Figure B.2 separates video length, the duration of an individual positive interval, and its normalized onset location. Video durations come from the media container metadata; event boundaries come from the retained annotations. Temporal distributions are normalized separately for each derivative, so different source sizes do not obscure distributional differences. The 8s and 30s cutoffs define the same instance-duration bins as Experiment 5. Normalized onset is s/T_v; it describes where events occur and is not an accuracy measure. These plots describe the exports without claiming uniform temporal coverage or an unbiased domain distribution.

B.3 Candidate Load and Rejection Targets

Figure B.3 shows how many unique descriptions a video requires the model to monitor, the balance of present and absent candidates, and the difficulty labels actually available. The benchmark tests absence of the specified event, not absence of all its entities. For example, a semantically nearby description can share objects and setting while differing in its defining action or relation. Any illustrated positive–negative pair must be drawn from the retained annotations and verified against the complete recording; a few example frames alone cannot establish whole-video absence.

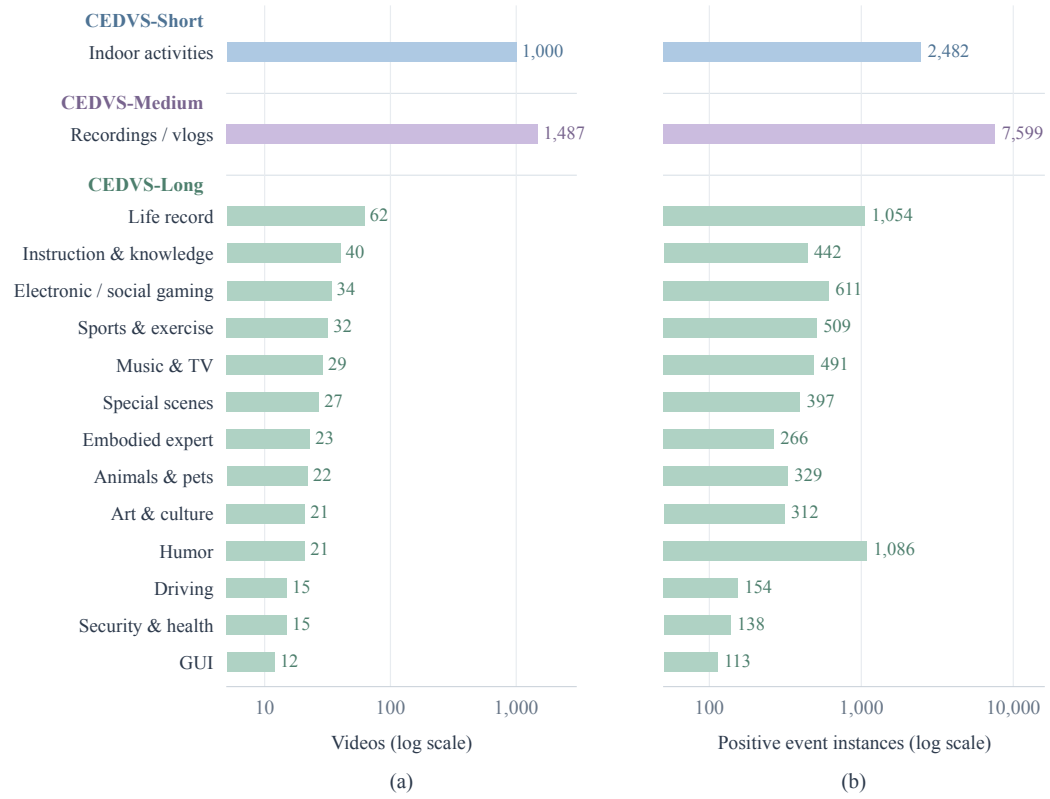

Figure B.1: Content-category coverage: (a) videos and (b) positive event instances in corresponding rows. CEDVS-Short/CEDVS-Medium use source-level content categories; CEDVS-Long uses its original detailed taxonomy. Logarithmic axes accommodate the different category sizes; counts are labeled directly.

B.4 Human Spot Checks

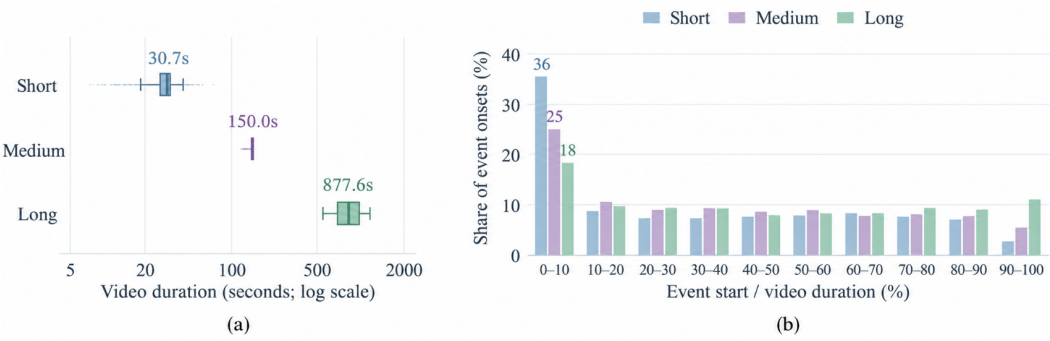

Figure B.2: Temporal characteristics of the three CEDVS-BENCH subsets. (a) Distribution of video duration for CEDVS-Short, CEDVS-Medium, and CEDVS-Long on a logarithmic scale; boxes indicate quartiles and medians, and whiskers extend to 1.5 times the interquartile range. The median durations are 30.7, 150.0, and 877.6 seconds, respectively. (b) Distribution of event onset positions normalized by video duration. Event starts are grouped into ten equal relative-time intervals from 0–10% to 90–100%, and the percentages for each subset sum to 100% before rounding. For example, 36% of CEDVS-Short event instances begin within the first 10% of the video. Dataset statistics are summarized in Table 2.

Audit scope and automatic filtering. Human inspection covered 110 videos across the three sources, comprising 524 positive records and 2,044 negative candidates before automatic filtering. Within this reviewed sample, filtering removed 64 negative candidates from QVHighlights and 18 positive records and 142 negative candidates from CGBench. No reviewed Charades-STA

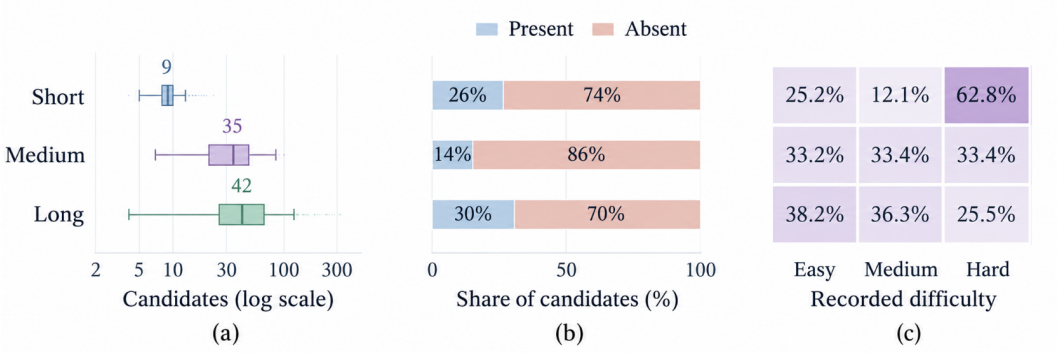

Figure B.3: Candidate composition of the three CEDVS-BENCH subsets. (a) Distribution of the number of unique event candidates per video on a logarithmic scale; boxes indicate quartiles and medians, with median candidate counts of 9, 35, and 42 for CEDVS-Short, CEDVS-Medium, and CEDVS-Long, respectively. (b) Relative proportions of present and absent candidates in each subset. Absent candidates account for 74%, 86%, and 70% of the candidates in CEDVS-Short, CEDVS-Medium, and CEDVS-Long, respectively. (c) Distribution of recorded difficulty levels among absent candidates, reported as Easy, Medium, and Hard. Difficulty memberships are not necessarily mutually exclusive, and percentages are rounded.

records were removed. The resulting retained audit sample contained 506 positive records and 1,838 negative candidates. Source-specific sample sizes and validity rates are reported in Table 3. This data-quality audit is separate from the model-behavior case analysis in Appendix H.

Validity criteria and denominators. Positive validity requires both event occurrence and a faithful description of the observed content. Negative validity requires the specified event to be absent throughout the video. Reviewers distinguished confirmed, rejected, and unresolved positive records, and valid, invalid, and unresolved negative candidates.

Among the retained positive records, 471 were confirmed, 34 were rejected, and one remained unresolved. Positive validity uses all retained positive records as its denominator. Among the retained negative candidates, 1,788 were valid, 42 were invalid, and eight remained unresolved. Negative validity uses valid plus invalid candidates as its denominator, excluding unresolved judgments. Unresolved counts are reported separately and are not treated as verified annotations. Automatic removal and human rejection are distinct: the filtering counts above do not imply that all human-rejected records were removed.

Negative difficulty groups. We additionally examine negative validity within the dataset's original Easy, Medium, and Hard categories. Across the retained audit sample, the corresponding rates were $640/651 = 98.31\%$, $552/569 = 97.01\%$, and $596/610 = 97.70\%$, respectively. The denominators exclude three unresolved Easy candidates and five unresolved Hard candidates; no Medium candidate remained unresolved. These statistics assess event absence within each category, rather than human agreement with the assigned difficulty labels.

Temporal boundary review. Boundary inspection assessed whether the annotated start and end times were acceptable and recorded revisions where needed. Among 483 records with determinate boundary judgments, 435 passed review and 48 did not. Three additional records had unresolved boundary judgments, while the remaining 20 retained positive records were not included in these boundary statistics.

Table B.1 reports boundary acceptance and the differences between original and reviewed timestamps. The tolerance-based rates require both the start and end times to differ by at most the stated tolerance.

Accepted boundaries were retained unchanged and therefore contributed zero difference. These measurements describe an accept-or-correct review procedure, rather than agreement between independently produced temporal annotations. Boundary-review eligibility is separate from positive-event validity: the temporal statistics are not restricted to the 471 confirmed positive records.

Table B.1: Temporal boundary-review results. Acceptance reports accepted annotations divided by records with determinate boundary judgments. The tolerance columns report the percentage of records for which both endpoints differ from the reviewed boundaries by at most the stated tolerance. Accepted boundaries remain unchanged and contribute zero difference.

Source	Acceptance count (%)	Both within 1 s (%)	Both within 2 s (%)	Both within 3 s (%)
CEDVS-Short	76/85 (89.41)	91.76	97.65	97.65
CEDVS-Medium	164/181 (90.61)	94.48	97.24	97.24
CEDVS-Long	195/217 (89.86)	96.31	99.54	100.00
Pooled	435/483 (90.06)	94.82	98.34	98.55

Scope of interpretation. The pooled statistics describe the retained audit sample and are not weighted estimates for the full benchmark. In particular, the CGBench audit covered only thirty videos despite containing many event records. Event counts therefore do not imply broad video-level coverage. Source-specific reporting makes these coverage limitations and differences in annotation quality explicit.

C Exact Interface, Matching, and Timing Definitions

C.1 Shared-Window Parsing and Candidate Identity

The following parser describes shared-window VLM runs; native adapters are specified in Appendix D. Frame selection uses source presentation timestamps. The decoder advances and caches required observations without providing imagery after any target time. Reading media duration or frame metadata in advance does not reveal future image content, but the implementation remains causal replay of a recording rather than an endpoint-discovering live interface.

Inactive candidates appear as A1, A2, etc., and active candidates as B1, B2, etc. The output syntax is illustrated below; it is not a verbatim reproduction of the full inference prompt. The exact saved prompt and raw responses should accompany reproducibility material.

```
START: [A1] | Frame: <frame_id>
END: [B1] | Frame: <frame_id>
NONE
```

Parsing uses the ID map frozen before the call. END is processed before START, but lists are not reinterpreted using later updates; a candidate that was inactive before the call cannot both start and end in that reply. The parser ignores unknown IDs and unmatched lines, clamps integer frame indices to the available range, and accepts at most one transition of each applicable type per candidate per response. END on the first or final sampled frame is ignored; START may occur on a boundary frame.

The implementation does not impose a complete cross-window requirement that a new START follow the preceding END. Main metrics use actually accepted onsets; interval-overlap diagnostics must separately document invalid intervals and any repairs. NONE and unparseable responses can both produce no update, so raw outputs are needed to distinguish deliberate silence from an interface failure. Initial state and subsequent transitions are predictions, never teacher-forced ground truth. Exact candidate grouping trims outer whitespace only; case and internal-space differences remain distinct. Identical intervals are deduplicated, while other overlapping intervals remain separate instances. Positive/negative label conflicts and overlapping difficulty memberships are recorded; difficulty counts therefore need not partition all negatives.

C.2 Lexicographic Matching Objective

For DT/SRT, let $a_r = d_r$ or c_r, respectively. Among one-to-one matches satisfying the relevant eligibility rule, the objective is lexicographic:

$$\mathcal{M}^* = \arg\min_{\mathcal{M}} \left(-|\mathcal{M}|, \sum_{(r,m)\in\mathcal{M}} (a_r - s_m) \right). \tag{C.1}$$

After sorting in temporal order, dynamic programming chooses among skipping a prediction, skipping a true instance, or matching an eligible pair. Empty prefixes have zero matches and zero cost. OED retains greedy nearest-onset matching, so an OED–DT gap cannot be interpreted as changing the time variable alone. We retain these implemented matching rules throughout; no alternative onset-matching experiment is included in the present ablation suite.

C.3 OEMD and Coverage

Let $\mathcal{H}$ contain every true onset whose positive candidate has at least one accepted START. The implemented conditional onset error and its coverage are

$$\text{OEMD} = \frac{1}{|\mathcal{H}|} \sum_{(v,i,s)\in\mathcal{H}} \min_{r\in\mathcal{R}_{vi}} |\hat{s}_r - s|, \qquad \text{Coverage} = \frac{|\mathcal{H}|}{\sum_{v,i} M_{vi}}. \tag{C.2}$$

A prediction may be nearest to multiple true onsets. Candidates with no START are excluded from OEMD, and an empty $\mathcal{H}$ gives an undefined mean. Coverage is not tolerance-based recall: one START includes every true onset of that positive candidate in $\mathcal{H}$, even if some occurrences remain unmatched. The evaluator stores the OEMD numerator count; coverage is derived by dividing that count by all retained true onsets. Reporting the conditional mean alone can favor a model that triggers on only a small subset of candidates.

C.4 Service Costs, Queues, and Response Lag

The recorded window service time includes preparation, model invocation, parsing, and state updates. Input serialization and pre-call request recording fall inside the current timer. Post-response result logging has a separate timer and is excluded. The first window may include media-metadata initialization. Model-call duration is saved separately and is not the entire service cost; loading and warm-up are excluded from the window replay.

For the grouped ablations in Appendix G, one window prepares and encodes its visual input once, then processes candidate groups sequentially. Let $E_{k,g}$ be measured elapsed service from the beginning of window k through completion of group g, excluding the separately timed result journals. A report emitted by that group is available at

$$c_r = B_k + E_{k,g(r)}, \qquad C_k = B_k + S_k, \qquad S_k = E_{k,g_{\text{last}}}. \tag{C.3}$$

The next window waits for the entire preceding window, while an earlier group's report need not wait for later groups. The one-group baseline is the special case $c_r = C_k$. The group order is seeded and rotated across seconds. All group calls contribute to window cost; the per-call mean is not substituted for total window service.

Queue waiting is $Q_k = B_k - d_k$ and full-window response lag is $L_k = C_k - d_k = Q_k + S_k$. Full-window lag summaries include all complete windows, including silent ones. In the ablation outputs this is `Full_Window_Response_Lag`, whereas `Response_Lag` is a distribution over emitted START reports and is not interchangeable with it. The pooled real-time factor is

$$\text{RTF}_{\text{pool}} = \frac{\sum_v \sum_k S_{vk}}{\sum_v \lfloor T_v \rfloor \times 1\text{s}}. \tag{C.4}$$

Under the complete one-second protocol this equals pooled mean window seconds divided by one second, not an unweighted average of per-video ratios. Changing stride requires an explicit observation-time denominator; window count can no longer stand in for video seconds. Conditional DT/SRT delays use their own matched sets and must retain match counts.

C.5 Exposure Normalization and Uncertainty

For $A_v = \sum_{i:y_{vi}=0} |\mathcal{R}_{vi}|$ negative STARTs and actual horizon $H_v = \lfloor T_v \rfloor$ seconds,

$$\mathrm{FA}_{\text{video-min}} = \frac{\sum_v A_v}{\sum_v H_v/60}, \qquad \mathrm{FA}_{\text{query-min}} = \frac{\sum_v A_v}{\sum_v n_v^- H_v/60}. \tag{C.5}$$

These rates are post-processing statistics derived from saved START reports and observation horizons; they are not additional inference conditions. Each absent candidate contributes once to the overall numerator even if it belongs to several difficulty groups. They accompany the fraction of absent candidates ever triggered. A wrongly activated candidate that never ends may generate only one START despite failing sustained rejection. Exposure normalization also does not remove differences in negative composition, domain, or prompt load.

Ablation uncertainty uses paired video-level resampling, preserving within-video event and window dependence and recomputing micro-averaged counts in each replicate. Table G.4 reports 95% percentile intervals from 2,000 replicates with seed 20260922. This is post-processing of saved outputs, not additional model inference; no intervals are claimed for tables without them.

D Model Configurations and Streaming Interfaces

D.1 Hardware and Inference Parameters

Experiments use NVIDIA A100 GPUs with 80 GB memory. Shared-window inference uses deterministic decoding (temperature 0, or greedy decoding), with one decision per video second and at most five causal frames sampled at 1 fps. No benchmark-specific training is performed. For CEDVS-Short/CEDVS-Medium, the maximum generated-token budget per call is 512 for Qwen3-VL, InternVL3.5, MiniCPM-V, VideoLLaMA3, LLaVA-OneVision, LLaVA-NeXT-Video, LongVA, and LongVU. CEDVS-Long budgets follow the corresponding saved configurations. The ablations retain temperature 0 and a 512-token limit per call. Service measurements include preparation, inference, parsing, and state updates, but exclude loading and warm-up; runtime load can still vary across executions. Native models require their own sampling, memory, and scheduling conventions rather than inheriting the five-frame restriction.

Native streaming interfaces and timing comparability. We evaluate Dispider, VideoLLM-Online, and M4 through their native streaming interfaces. Each model monitors one candidate event per run, maintains its streaming state throughout the video, and resets that state before replaying the video for the next candidate. This differs from the shared-window setting, where all candidates are evaluated jointly at each observation step.

Dispider processes 16-frame decision windows spanning 16 seconds (approximately 1 FPS), while retaining its decision KV cache and streaming history. It makes one native decision per window; emitted trigger timestamps serve as both predicted onsets and decision times. VideoLLM-Online receives one frame every 0.5 seconds and maintains its streaming state and KV cache. The monitored event is specified in the system prompt. Native inference is invoked after each incoming frame, and a nonempty response is treated as a trigger, with the corresponding video timestamp used as both the predicted onset and decision time. M4 receives one frame per second, maintains its streaming KV cache and history, and uses attention over historical observations for proactive grounding. We use its `REQUIREMENT MEET AT` timestamp as the predicted onset and `Current Time` as the decision time.

Response-time measurements are not uniform across these interfaces. For VideoLLM-Online and M4, the response timestamp is estimated by adding the recorded current-step processing cost to the decision timestamp. This cost includes visual processing, state updates, and text generation, as well as grounding for M4, but excludes offline video preprocessing; VideoLLM-Online also excludes advance video loading. These estimates are evaluated with a three-second response tolerance. They do not accumulate unfinished work from preceding steps and therefore differ from the FIFO service-cost replay used for shared-window SRT.

Accordingly, the native results assess task compatibility and provide a preliminary comparison under the specified adaptations. Differences in observation cadence, candidate batching, state main-

tenance, and timing instrumentation prevent a controlled comparison of response efficiency across model families. We therefore do not use the native response scores for a joint ranking with shared-window SRT, nor interpret the reported performance differences as architecture-level advantages or limitations.

E Complete Results and Efficiency

E.1 Complete Detection Scores

Table E.1 reports Precision/Recall/F1 at 1, 2, and 3 seconds for the full corresponding CEDVS subsets, consistent with Table 4. Native streaming VLMs are not included.

Table E.1: Complete detection scores (%) on the full corresponding CEDVS-Bench. P/R/F1 denote precision, recall, and F1. Tolerances are in seconds; coverage follows Table 4.

Model	Criterion	P@1	R@1	F1@1	P@2	R@2	F1@2	P@3	R@3	F1@3
CEDVS-Short										
Qwen3-VL-8B	Onset	23.12	42.75	30.01	28.55	52.78	37.05	32.16	59.47	41.75
	Decision	20.70	38.28	26.87	27.48	50.81	35.67	30.07	55.60	39.03
	Response	1.44	2.66	1.87	4.47	8.26	5.80	11.27	20.83	14.62
InternVL3.5-8B	Onset	12.38	39.52	18.85	13.90	44.40	21.17	15.23	48.63	23.19
	Decision	9.70	30.98	14.78	12.94	41.34	19.71	14.07	44.92	21.42
	Response	0.10	0.32	0.15	0.69	2.22	1.06	3.53	11.28	5.38
MiniCPM-V-2.6	Onset	12.19	32.47	17.72	14.61	38.92	21.24	16.65	44.36	24.21
	Decision	10.51	28.00	15.28	12.01	31.99	17.46	13.34	35.54	19.40
	Response	0.02	0.04	0.02	0.26	0.68	0.37	2.69	7.17	3.91
LLaVA-OneVision-1.5-8B	Onset	10.95	38.28	17.02	13.19	46.13	20.52	15.41	53.87	23.96
	Decision	9.71	33.96	15.11	12.84	44.88	19.96	14.21	49.68	22.09
	Response	0.01	0.04	0.02	0.08	0.28	0.13	1.06	3.71	1.65
LLaVA-NeXT-Video-7B	Onset	5.65	35.62	9.76	6.41	40.37	11.06	7.04	44.32	12.14
	Decision	5.48	34.49	9.45	6.02	37.91	10.39	6.29	39.65	10.86
	Response	0.00	0.00	0.00	0.01	0.08	0.02	1.06	6.65	1.82
VideoLLaMA3-7B	Onset	10.49	34.62	16.10	12.62	41.66	19.37	14.39	47.51	22.09
	Decision	7.63	25.18	11.71	10.89	35.95	16.72	12.30	40.61	18.88
	Response	0.00	0.00	0.00	0.08	0.27	0.12	1.83	6.04	2.81
LongVA-7B	Onset	6.86	36.46	11.55	8.18	43.47	13.77	9.35	49.68	15.73
	Decision	5.96	31.67	10.03	6.96	36.99	11.71	7.42	39.44	12.49
	Response	0.00	0.00	0.00	0.06	0.32	0.10	0.78	4.15	1.31
LongVU-7B	Onset	18.08	35.66	23.99	21.47	42.34	28.49	24.68	48.67	32.75
	Decision	16.47	32.47	21.85	19.37	38.20	25.70	20.90	41.22	27.73
	Response	0.00	0.00	0.00	0.10	0.20	0.14	3.21	6.33	4.26
CEDVS-Medium										
Qwen3-VL-8B	Onset	7.33	20.88	10.85	10.08	28.73	14.92	13.43	38.27	19.88
	Decision	8.01	22.83	11.86	12.46	35.50	18.44	14.22	40.52	21.05
	Response	0.98	2.80	1.46	3.77	10.75	5.58	6.58	18.77	9.75
InternVL3.5-8B	Onset	3.06	17.98	5.23	3.91	23.02	6.69	4.66	27.40	7.96
	Decision	2.58	15.19	4.41	3.95	23.25	6.76	4.53	26.62	7.74
	Response	0.04	0.25	0.07	0.18	1.07	0.31	0.39	2.30	0.67
MiniCPM-V-2.6	Onset	2.60	20.11	4.61	3.39	26.23	6.01	3.82	29.49	6.76
	Decision	1.90	14.69	3.37	2.50	19.29	4.42	2.90	22.40	5.13
	Response	< 0.01	0.01	< 0.01	0.01	0.07	0.02	0.02	0.17	0.04
LLaVA-OneVision-1.5-8B	Onset	3.54	18.72	5.95	4.81	25.45	8.09	5.99	31.70	10.07
	Decision	3.05	16.16	5.14	5.04	26.65	8.47	5.91	31.23	9.94
	Response	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.11	0.03
LLaVA-NeXT-Video-7B	Onset	2.16	16.24	3.81	2.32	17.45	4.09	2.39	18.00	4.22
	Decision	2.11	15.90	3.73	2.22	16.71	3.92	2.26	17.04	4.00
	Response	0.00	0.00	0.00	0.00	0.00	0.00	< 0.01	0.01	< 0.01
VideoLLaMA3-7B	Onset	2.39	17.49	4.20	2.82	20.63	4.96	3.00	22.02	5.29
	Decision	1.37	10.03	2.41	2.09	15.29	3.67	2.50	18.34	4.41
	Response	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Continued on next page

Table E.1 (continued)

Model	Criterion	P@1	R@1	F1@1	P@2	R@2	F1@2	P@3	R@3	F1@3
LongVA-7B	Onset	2.81	16.44	4.80	3.57	20.90	6.11	4.14	24.21	7.07
	Decision	2.35	13.74	4.01	3.01	17.58	5.14	3.45	20.16	5.89
	Response	0.00	0.00	0.00	0.00	0.00	0.00	< 0.01	0.01	< 0.01
LongVU-7B	Onset	4.49	14.70	6.88	5.91	19.36	9.06	6.69	21.88	10.24
	Decision	3.41	11.15	5.22	4.39	14.36	6.72	5.22	17.09	8.00
	Response	0.00	0.00	0.00	0.00	0.00	0.00	< 0.01	0.01	0.01
CEDVS-Long										
Qwen3-VL-8B	Onset	3.21	10.25	4.88	4.21	13.45	6.41	5.11	16.33	7.78
	Decision	2.72	8.69	4.14	3.96	12.68	6.04	4.60	14.69	7.00
	Response	0.00	0.00	0.00	0.01	0.04	0.02	0.04	0.12	0.06
InternVL3.5-8B	Onset	1.84	5.48	2.75	2.24	6.68	3.36	2.62	7.80	3.92
	Decision	1.01	3.00	1.51	1.72	5.14	2.58	2.12	6.33	3.18
	Response	0.02	0.05	0.03	0.02	0.07	0.03	0.03	0.08	0.04
MiniCPM-V-2.6	Onset	1.03	4.48	1.68	1.35	5.86	2.20	1.54	6.69	2.51
	Decision	0.88	3.81	1.43	1.02	4.42	1.66	1.12	4.85	1.82
	Response	0.01	0.05	0.02	0.01	0.05	0.02	0.02	0.08	0.03
LLaVA-OneVision-1.5-8B	Onset	1.58	6.11	2.51	1.98	7.65	3.15	2.34	9.04	3.72
	Decision	1.08	4.17	1.72	1.72	6.63	2.73	2.11	8.14	3.35
	Response	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLaVA-NeXT-Video-7B	Onset	1.54	5.73	2.43	1.60	5.95	2.52	1.64	6.10	2.59
	Decision	1.53	5.69	2.42	1.58	5.86	2.49	1.59	5.90	2.50
	Response	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VideoLLaMA3-7B	Onset	1.44	7.14	2.40	1.72	8.52	2.86	1.86	9.23	3.10
	Decision	1.17	5.82	1.95	1.35	6.68	2.24	1.51	7.48	2.51
	Response	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LongVA-7B	Onset	4.08	3.55	3.80	4.53	3.95	4.22	4.89	4.26	4.55
	Decision	3.55	3.09	3.30	4.04	3.52	3.76	4.32	3.76	4.02
	Response	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LongVU-7B	Onset	1.94	3.84	2.57	2.22	4.40	2.95	2.51	4.97	3.33
	Decision	1.37	2.72	1.82	1.65	3.27	2.19	2.03	4.03	2.70
	Response	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

E.2 Processing Costs and Response Delay

Service statistics include every completed window, including silent decisions. Mean and P95 service are available in every run summary; conditional decision/response delays use their own matches and retain match counts. Under the one-second cadence, pooled RTF numerically equals mean service in seconds.

Table E.2: Timing in seconds; RTF is dimensionless. Final columns are conditional mean delay / matched count at 3s. All columns use the corresponding main-table cohort.

Model	Mean service	P95 service	RTF	Decision delay / n	Response delay / n
CEDVS-Short					
Qwen3-VL-8B	1.36	2.80	1.36	1.09 / 1380	2.01 / 517
InternVL3.5-8B	1.47	2.38	1.47	1.14 / 1115	2.33 / 280
MiniCPM-V-2.6	2.23	4.52	2.23	1.11 / 882	2.48 / 178
LLaVA-OneVision-1.5-8B	2.94	5.89	2.94	1.15 / 1233	2.63 / 92
LLaVA-NeXT-Video-7B	1.85	6.29	1.85	1.00 / 984	2.69 / 165
VideoLLaMA3-7B	3.25	7.39	3.25	1.27 / 1008	2.71 / 150
LongVA-7B	1.91	4.50	1.91	1.11 / 979	2.59 / 103
LongVU-7B	2.39	3.57	2.39	1.13 / 1023	2.69 / 157
CEDVS-Medium					
Qwen3-VL-8B	0.68	1.07	0.68	1.43 / 3079	1.91 / 1426
InternVL3.5-8B	1.10	1.72	1.10	1.44 / 2023	2.04 / 175
MiniCPM-V-2.6	2.31	8.33	2.31	1.41 / 1702	2.20 / 13
LLaVA-OneVision-1.5-8B	1.55	2.35	1.55	1.53 / 2373	2.54 / 8
LLaVA-NeXT-Video-7B	1.28	5.29	1.28	1.06 / 1295	2.52 / 1

Continued on next page

Table E.2 (continued)

Model	Mean service	P95 service	RTF	Decision delay / n	Response delay / n
VideoLLaMA3-7B	2.48	4.07	2.48	1.56 / 1394	– / 0
LongVA-7B	1.84	3.90	1.84	1.40 / 1532	2.80 / 1
LongVU-7B	2.05	3.20	2.05	1.44 / 1299	2.99 / 1
CEDVS-Long					
Qwen3-VL-8B	1.37	2.30	1.37	1.43 / 837	2.43 / 7
InternVL3.5-8B	0.95	1.40	0.95	1.61 / 373	1.47 / 5
MiniCPM-V-2.6	2.04	6.23	2.04	1.23 / 286	1.39 / 5
LLaVA-OneVision-1.5-8B	2.78	5.87	2.78	1.57 / 480	– / 0
LLaVA-NeXT-Video-7B	1.61	3.00	1.61	1.01 / 347	– / 0
VideoLLaMA3-7B	1.88	3.67	1.88	1.27 / 441	– / 0
LongVA-7B	2.68	4.70	2.68	1.22 / 219	– / 0
LongVU-7B	2.53	3.87	2.53	1.48 / 234	– / 0

E.3 Negative Rejection across Sources

Table E.3 reports positive onset recall, overall and difficulty-specific FAR, and exposure-normalized false STARTs. CEDVS-Short difficulty metadata are joined from the supplied Easy/Medium export only after verifying identical video/candidate identities and positive labels. Generated-hard-negative tags provide Hard membership. This join changes no model input or prediction; one CEDVS-Short absent candidate belongs to both Medium and Hard. Each absent candidate is counted once in overall FAR. The contrast is not simply between high and low recall: on CEDVS-Long, LongVA has relatively low overall FAR (18.92%) but low positive recall, whereas several models trigger on most absent candidates. Reliable monitoring requires both detection and sustained rejection.

Table E.3: Negative rejection on the main-table cohorts. Recall and FAR are percentages; alarm rates count negative STARTs, including repeats. Difficulty memberships may overlap.

Model	Recall	Overall FAR	Easy FAR	Medium FAR	Hard FAR	FA / video-min	FA / query-min
CEDVS-Short							
Qwen3-VL-8B	59.47	23.90	5.65	33.85	29.29	3.782	0.545
InternVL3.5-8B	48.63	78.29	65.20	85.23	82.21	11.226	1.619
MiniCPM-V-2.6	44.36	52.40	31.45	52.22	60.83	8.263	1.191
LLaVA-OneVision-1.5-8B	53.87	72.34	54.59	81.63	77.67	11.442	1.650
LLaVA-NeXT-Video-7B	44.32	100.00	100.00	100.00	100.00	22.143	3.193
VideoLLaMA3-7B	47.51	80.22	72.00	80.74	83.19	11.447	1.637
LongVA-7B	49.68	93.93	90.94	92.32	95.44	18.117	2.612
LongVU-7B	48.67	42.54	18.87	34.45	53.61	6.093	0.879
CEDVS-Medium							
Qwen3-VL-8B	38.27	30.05	13.38	31.62	45.04	3.814	0.125
InternVL3.5-8B	27.40	82.17	67.91	86.07	92.44	10.084	0.330
MiniCPM-V-2.6	29.49	81.21	76.46	81.68	85.47	12.771	0.417
LLaVA-OneVision-1.5-8B	31.70	70.12	53.97	73.55	82.74	8.808	0.288
LLaVA-NeXT-Video-7B	18.00	100.00	100.00	100.00	99.99	13.021	0.426
VideoLLaMA3-7B	22.02	98.44	98.23	98.49	98.59	12.716	0.416
LongVA-7B	24.21	68.59	67.04	67.28	71.45	9.693	0.317
LongVU-7B	21.88	44.61	40.95	42.48	50.38	5.484	0.179
CEDVS-Long							
Qwen3-VL-8B	16.33	68.16	39.56	80.64	93.39	2.072	0.051
InternVL3.5-8B	7.80	85.92	67.38	96.35	98.86	2.274	0.056
MiniCPM-V-2.6	6.69	81.59	77.53	83.03	85.58	3.118	0.079
LLaVA-OneVision-1.5-8B	9.04	81.77	61.04	93.02	96.81	2.810	0.069
LLaVA-NeXT-Video-7B	6.10	97.38	97.77	96.83	97.57	2.928	0.072
VideoLLaMA3-7B	9.23	99.83	99.70	99.92	99.89	3.677	0.090
LongVA-7B	4.26	18.92	17.31	17.88	22.86	0.640	0.016
LongVU-7B	4.97	58.97	54.07	60.47	64.18	1.575	0.038

E.4 Offline Presence and Output Coverage

Table E.4 complements the same four-model, matched-video comparison in Table 5. Presence precision/recall use unique candidate descriptions, with positive labels winning conflicts and distinct true intervals retained. Output coverage is video-level: a completed answer contains at least one parsed event interval or an explicit candidate-tagged NONE record. This observable check does not certify correctness, complete candidate enumeration, or every output line. Both conditions retain their own coverage. Malformed completed outputs remain in score denominators as no detections; explicit execution failures are excluded from every corresponding condition.

Table E.4: Offline candidate-presence scores and video output coverage (%). Coverage is the fraction with a parsed interval or candidate-tagged NONE, not semantic accuracy.

Model	Presence precision	Presence recall	Presence F1	Output coverage	Timestamp coverage
CEDVS-Short					
Qwen3-VL-8B	80.77	76.32	78.48	100.00	100.00
InternVL3.5-8B	81.64	54.11	65.08	100.00	100.00
MiniCPM-V-2.6	65.04	36.10	46.43	49.60	31.10
VideoLLaMA3-7B	61.51	11.58	19.49	20.38	18.15
CEDVS-Medium					
Qwen3-VL-8B	45.42	75.16	56.63	100.00	98.18
InternVL3.5-8B	60.21	65.54	62.76	99.46	99.46
MiniCPM-V-2.6	25.05	11.18	15.46	37.53	21.59
VideoLLaMA3-7B	23.91	14.87	18.34	36.99	25.62

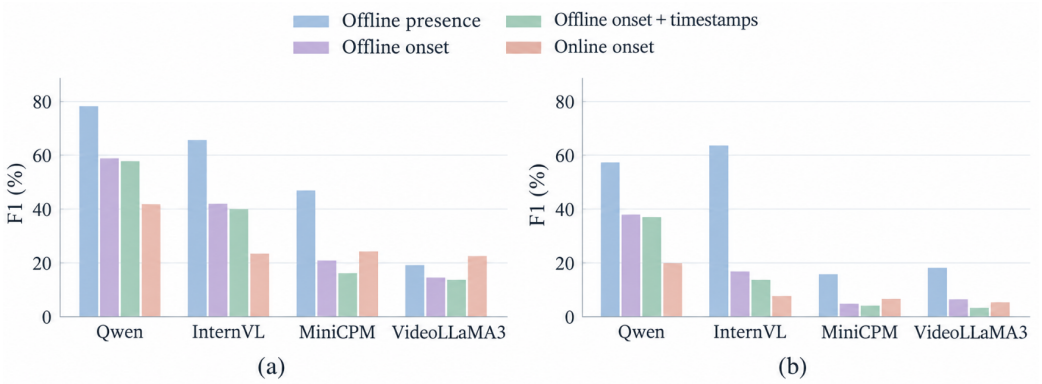

Figure E.1: Comparison of offline recognition, offline localization, and online onset detection on matched four-model cohorts. (a) Results on CEDVS-Short. (b) Results on CEDVS-Medium. *Offline presence* measures whether the target event is recognized in the full video, while *offline onset* evaluates temporal onset localization from the same offline output. *Offline onset + timestamps* additionally provides explicit frame timestamps during inference, whereas *online onset* uses causal recent-window observations. The results illustrate that strong offline recognition does not necessarily translate into accurate temporal localization or online event detection.

F Event Duration and Relative Occupancy

F.1 Instance-Duration Analysis

Short, Medium, and Long mean $0 < D \leq 8$, $8 < D \leq 30$, and $D > 30$ seconds. Predictions are matched globally before true instances are assigned to bins; complete misses remain in recall denominators. Tables F.1 and F.2 use each model's full completed cohort from Table 4. No false positives are arbitrarily assigned to true-duration bins.

Table F.1: Instance-duration recall (%) at 3s. S/M/L denote the fixed Short/Medium/Long bins. Onset and decision matching use their respective rules.

Model	Onset S	Onset M	Onset L	Decision S	Decision M	Decision L
CEDVS-Short						
Qwen3-VL-8B	52.26	65.57	90.91	49.72	60.62	80.30
InternVL3.5-8B	38.48	57.24	92.42	36.03	52.56	81.82
MiniCPM-V-2.6	39.43	48.92	59.09	31.99	38.42	53.03
LLaVA-OneVision-1.5-8B	45.05	61.23	93.94	42.44	55.42	87.88
LLaVA-NeXT-Video-7B	34.44	52.04	98.48	29.14	48.14	92.42
VideoLLaMA3-7B	37.18	56.28	88.89	32.40	47.49	75.00
LongVA-7B	40.78	57.07	90.91	30.64	46.92	77.27
LongVU-7B	39.35	56.46	90.91	33.10	48.05	77.27
CEDVS-Medium						
Qwen3-VL-8B	41.54	37.35	37.31	46.17	38.62	39.45
InternVL3.5-8B	23.44	26.19	32.90	24.33	25.18	31.20
MiniCPM-V-2.6	31.04	29.61	28.00	22.61	22.53	21.98
LLaVA-OneVision-1.5-8B	27.74	31.54	35.23	28.15	30.66	34.98
LLaVA-NeXT-Video-7B	11.75	18.14	22.85	11.22	16.90	22.08
VideoLLaMA3-7B	16.85	22.32	25.67	13.59	18.32	22.27
LongVA-7B	21.01	23.96	27.32	15.73	19.60	24.84
LongVU-7B	18.69	20.40	27.27	13.71	15.65	22.56
CEDVS-Long						
Qwen3-VL-8B	17.12	17.45	14.95	18.43	16.21	11.83
InternVL3.5-8B	8.20	8.29	7.40	6.70	6.50	6.00
MiniCPM-V-2.6	8.32	6.48	6.31	5.03	4.76	4.87
LLaVA-OneVision-1.5-8B	8.45	9.62	8.72	9.44	8.71	7.06
LLaVA-NeXT-Video-7B	4.19	6.68	6.31	4.09	6.35	6.19
VideoLLaMA3-7B	11.43	10.32	7.27	7.75	8.58	6.30
LongVA-7B	4.09	4.32	4.26	3.49	3.77	3.85
LongVU-7B	4.25	5.00	5.23	2.73	4.12	4.46

Table F.2: Conditional onset error (seconds) / coverage (%) by duration. Coverage counts true onsets whose candidate has any accepted START, not correctly matched onsets.

Model	Short	Medium	Long
CEDVS-Short			
Qwen3-VL-8B	5.33 / 83.06	3.18 / 90.55	0.80 / 96.97
InternVL3.5-8B	9.10 / 98.89	5.12 / 99.83	0.82 / 100.00
MiniCPM-V-2.6	7.47 / 80.44	4.87 / 79.97	1.39 / 68.18
LLaVA-OneVision-1.5-8B	7.82 / 98.42	4.34 / 99.65	0.72 / 100.00
LLaVA-NeXT-Video-7B	11.23 / 100.00	6.07 / 100.00	0.44 / 100.00
VideoLLaMA3-7B	8.53 / 90.70	4.66 / 96.93	1.09 / 100.00
LongVA-7B	8.74 / 97.31	5.09 / 97.75	1.25 / 100.00
LongVU-7B	5.83 / 72.29	3.69 / 84.30	0.66 / 96.97
CEDVS-Medium			
Qwen3-VL-8B	24.25 / 91.75	25.78 / 91.93	21.11 / 92.92
InternVL3.5-8B	44.42 / 99.53	41.00 / 99.58	31.20 / 99.66
MiniCPM-V-2.6	32.83 / 93.35	33.09 / 93.95	29.38 / 92.77
LLaVA-OneVision-1.5-8B	34.14 / 97.50	32.65 / 98.31	24.91 / 98.59
LLaVA-NeXT-Video-7B	61.34 / 100.00	56.29 / 100.00	43.08 / 100.00
VideoLLaMA3-7B	49.84 / 99.17	46.73 / 99.48	38.07 / 99.85
LongVA-7B	30.18 / 76.50	29.30 / 82.46	26.52 / 88.16
LongVU-7B	28.64 / 59.76	29.23 / 60.03	21.66 / 66.52
CEDVS-Long			
Qwen3-VL-8B	255.20 / 99.45	267.91 / 99.02	266.42 / 99.43
InternVL3.5-8B	336.8/ 99.10	373.7/ 99.59	359.9 / 99.68
MiniCPM-V-2.6	261.66 / 82.50	279.38 / 82.02	278.96 / 90.50
LLaVA-OneVision-1.5-8B	283.40 / 97.32	316.77 / 98.26	300.49 / 99.11
LLaVA-NeXT-Video-7B	360.90 / 94.42	403.42 / 94.48	386.06 / 99.27

Continued on next page

Table F.2 (continued)

Model	Short	Medium	Long
VideoLLaMA3-7B	294.46 / 100.00	338.34 / 99.92	342.69 / 99.88
LongVA-7B	233.80 / 15.27	224.59 / 18.19	213.99 / 29.93
LongVU-7B	226.95 / 53.44	271.05 / 56.28	278.73 / 68.64

Interpretation. High conditional coverage can coexist with poor temporal localization. On the common CEDVS-Long cohort, Qwen and InternVL cover more than 99% of true instances in every duration bin, yet onset recall remains below 18% and 9%, respectively. Their conditional errors are hundreds of seconds. One distant START can place all repeated occurrences of a candidate in the coverage denominator; it does not establish detection of each occurrence. Conversely, CEDVS-Short Long instances nearly fill the video and are few in number (66), so their high recall should not be generalized to long events in arbitrary streams.

F.2 Relative Occupancy within Duration Bins

Table F.3 uses the same matched video sets. Within each source and duration bin, the annotation median of D/T_v defines the split; ties enter the lower group. Both models share boundaries and instance counts. Occupancy is observational and covaries with onset position, semantics, and candidate load. The differences in Figure 3(c) therefore describe associations, not an isolated occupancy effect.

Table F.3: Within-bin occupancy diagnostic. Median occupancy and recalls are percentages. Recall cells are Qwen / InternVL; both use identical instances.

Duration	Group	Median occupancy	Instances	Onset recall@3s	Decision recall@3s
CEDVS-Short					
Short	Lower	15.28	632	46.68 / 34.18	47.31 / 32.12
Short	Higher	15.28	631	57.84 / 42.79	52.14 / 39.94
Medium	Lower	47.03	577	57.54 / 46.62	52.51 / 42.81
Medium	Higher	47.03	576	73.61 / 67.88	68.75 / 62.33
Long	Lower	98.03	33	84.85 / 90.91	69.70 / 72.73
Long	Higher	98.03	33	96.97 / 93.94	90.91 / 90.91
CEDVS-Medium					
Short	Lower	4.00	943	39.13 / 23.33	48.25 / 24.71
Short	Higher	4.00	742	44.61 / 23.58	43.53 / 23.85
Medium	Lower	10.67	1993	39.09 / 27.04	39.49 / 25.89
Medium	Higher	10.67	1860	35.48 / 25.27	37.69 / 24.41
Long	Lower	33.33	1050	34.10 / 27.52	36.57 / 25.52
Long	Higher	33.33	1011	40.65 / 38.48	42.43 / 37.09
CEDVS-Long					
Short	Lower	0.69	463	16.41 / 7.34	21.38 / 6.26
Short	Higher	0.69	454	17.84 / 9.03	15.42 / 7.05
Medium	Lower	1.76	1172	18.52 / 7.59	17.66 / 6.23
Medium	Higher	1.76	1172	16.38 / 9.04	14.76 / 6.74
Long	Lower	7.77	1218	16.34 / 7.47	12.64 / 5.91
Long	Higher	7.77	1217	13.56 / 7.23	11.01 / 6.08

G Ablations of Visual Context and Candidate Grouping

G.1 Shared Design and Fixed Nested Subsets

We evaluate Qwen3-VL-8B and InternVL3.5-8B on CEDVS-Short and CEDVS-Medium. Each source has a fixed 100-video subset for recent-frame and sampled-history comparisons. A fixed 20-video subset nested within those 100 videos is used for candidate grouping. Both models use identical video identities in every corresponding condition. Within each source, videos are stratified by annotation duration and unique-candidate count; sample counts are allocated proportionally using

the largest-remainder rule, followed by seeded random sampling within strata. Duration cutoffs are 20/35 seconds for CEDVS-Short and 130 seconds for CEDVS-Medium; candidate-count strata use source-level tercile boundaries. Neither model scores nor completion status enters selection. Actual inference uses media duration rather than rounded annotation duration.

All conditions retain decisions at integer seconds starting from $t = 1$, the original parser, predicted active/inactive state, candidate membership, temperature zero, and a 512-token output limit per call. The three-video timing pilot for each source is nested within its 20-video group subset; complete pilot results are reused once and are not a separate reported experiment. Five-frame baseline scores are re-aggregated from compatible complete main-experiment outputs on exactly the same 100 or 20 videos. Compatible completed ten-frame outputs are also reused. No test-set tuning of candidate membership is performed.

Tables G.1–G.3 report the completed matched-subset experiments. Scores at a shared tolerance use the original matching rules. Overall FAR counts each unique absent candidate once; missing difficulty metadata do not become Hard labels. The 20-video grouping analysis is exploratory and should not support broad claims from small score differences. Historical baseline service measurements may reflect different server load or hardware conditions: accuracy comparisons use aligned annotations, but service time and cost-dependent Response F1 must retain their measurement provenance rather than be interpreted as strictly controlled speedups.

G.2 Recent-Frame Budget

With a maximum of $W \in \{5, 10, 20, 30\}$ recent frames, target source times at decision t are

$$\mathcal{U}_t^{(W)} = \{\max(0, t - W + 1), \ldots, t\}. \tag{G.1}$$

The latest source frame no later than each target is used. Thus the steady-state spans are $W - 1$ seconds at one frame per second; insufficient initial history is not padded with future frames. All candidates are evaluated together. This comparison tests the joint effect of a longer recent context and its increased visual input budget at a fixed sampling density and decision cadence. It does not isolate temporal span from frame count. We examine whether additional recent evidence changes onset localization, timely decisions, rejection, and total processing cost, without presupposing monotonic improvement.

Table G.1: Recent-frame ablation on the same 100 videos per source. Scores and FAR are percentages. Five frames is reaggregated from the corresponding main-run videos; service is measured whole-window time.

Model	Condition	Onset F1@3s	Decision F1@3s	Response F1@3s	FAR	Service (s)
CEDVS-Short						
Qwen3-VL-8B	Recent 5 frames	44.41	40.40	12.89	23.50	1.338
	Recent 10 frames	34.53	30.01	11.72	29.34	1.639
	Recent 20 frames	24.46	18.37	6.21	36.23	2.706
	Recent 30 frames	19.06	15.85	5.67	39.22	2.992
InternVL3.5-8B	Recent 5 frames	22.00	21.61	4.91	76.20	1.474
	Recent 10 frames	23.60	21.26	9.91	76.95	1.203
	Recent 20 frames	22.87	19.46	8.43	70.06	1.818
	Recent 30 frames	21.10	18.96	6.60	65.57	2.322
CEDVS-Medium						
Qwen3-VL-8B	Recent 5 frames	18.16	20.35	9.63	30.06	0.688
	Recent 10 frames	11.77	15.97	5.84	35.54	0.667
	Recent 20 frames	9.28	9.38	0.70	53.78	1.086
	Recent 30 frames	7.20	5.63	0.30	69.81	1.445
InternVL3.5-8B	Recent 5 frames	7.76	6.68	0.74	81.30	1.136
	Recent 10 frames	8.16	6.52	1.41	82.02	0.968
	Recent 20 frames	7.81	5.28	0.00	80.20	1.926
	Recent 30 frames	7.75	5.44	0.20	78.19	2.489

Results. Additional recent frames do not consistently improve monitoring. Qwen onset F1 decreases from 44.41 to 19.06 on CEDVS-Short and from 18.16 to 7.20 on CEDVS-Medium as the

cap grows from 5 to 30 frames; FAR rises from 23.50% to 39.22% and from 30.06% to 69.81%. InternVL is less sensitive in onset F1 (22.00 to 21.10; 7.76 to 7.75), and CEDVS-Short FAR decreases from 76.20% to 65.57%. More visible context therefore changes detection and rejection differently across models. The experiment jointly changes span and frame budget, and does not identify a unique mechanism for these changes.

G.3 Uniformly Sampled Causal History

The history condition draws at most 32 frames from the entire visible prefix, replacing the recent-window input. Its target times are

$$\mathcal{U}_t^{\text{hist}} = \begin{cases} \{0, 1, \ldots, t\}, & t \le 31, \\ \{\lfloor jt/31 \rfloor : j = 0, \ldots, 31\}, & t > 31. \end{cases} \tag{G.2}$$

The endpoints are included, and target times map to causal source-frame presentation timestamps. The prompt and output parser use these source times, not the artificial positions in the uploaded one-frame-per-second MP4. Historical integer-second observations are retained in the client cache to support changing sample positions without substituting later frames; at most 32 selected frames are submitted to the model at a decision. This is sampled visual history, not all one-frame-per-second observations, not 32 additional frames appended to a recent window, and not a learned recurrent memory.

The same predicted event state is retained in the reference and history conditions. Broadening coverage sacrifices temporal density as the prefix grows, and also changes the frame budget relative to the five-frame reference. The 30-frame recent condition provides a nearby-budget context, but is not an exactly frame-matched control. Interpret any difference as a comparison of these input policies, not an isolated causal effect of memory. Table G.2 examines whether dispersed historical evidence changes detection and rejection sufficiently to offset its cost and reduced local density.

Table G.2: Sampled causal history on the same 100 videos per source. History contains at most 32 frames in total. The five-frame reference is identical to Table G.1.

Model	Condition	Onset F1@3s	Decision F1@3s	Response F1@3s	FAR	Service (s)
CEDVS-Short						
Qwen3-VL-8B	Recent 5 frames	44.41	40.40	12.89	23.50	1.338
	Uniform history ≤ 32	19.20	15.97	5.61	38.02	2.850
InternVL3.5-8B	Recent 5 frames	22.00	21.61	4.91	76.20	1.474
	Uniform history ≤ 32	21.34	19.02	7.11	65.72	2.183
CEDVS-Medium						
Qwen3-VL-8B	Recent 5 frames	18.16	20.35	9.63	30.06	0.688
	Uniform history ≤ 32	1.96	1.40	0.05	95.23	2.479
InternVL3.5-8B	Recent 5 frames	7.76	6.68	0.74	81.30	1.136
	Uniform history ≤ 32	5.35	3.81	0.21	81.47	2.739

Results. Uniform history does not recover the five-frame baseline. Qwen falls to 19.20 onset F1 on CEDVS-Short and 1.96 on CEDVS-Medium, with CEDVS-Medium FAR reaching 95.23%. InternVL changes from 22.00 to 21.34 on CEDVS-Short and from 7.76 to 5.35 on CEDVS-Medium. On shorter CEDVS-Short videos, history and recent-30 inputs cover similar evidence and produce similar scores; on CEDVS-Medium, the history policy distributes a fixed budget across a longer prefix. These results support retaining temporal density as an explicit design variable, without attributing the gap solely to memory capacity.

G.4 Candidate Grouping at a Fixed Visual Input

On each source's fixed 20-video subset, we compare all candidates in one call, one candidate per call, and relation-based groups. A relation-based group contains a positive candidate and its associated retained negative descriptions; it need not contain exactly two descriptions. Candidates without a retained association remain standalone, and shared negatives are assigned once to the first eligible

positive anchor in the canonical candidate order. Exact-text grouping and positive-wins conflict handling remain unchanged. Ground-truth labels are used only to form this diagnostic partition and to score outputs, never shown in the inference prompt. The supplied partition is therefore an evaluation control rather than a grouping procedure available for arbitrary unlabeled deployment.

Every condition monitors the same complete set of unique descriptions, so this experiment varies joint prompt load rather than removing difficult candidates. The pair-group order is shuffled, and the seeded call order rotates across decision times. At each timestamp, recent five-frame video preparation, MP4 encoding, and base64 serialization occur once and are shared across groups; the model service receives the same visual payload for each group. This client-side reuse does not establish that the server computes visual features only once. All API calls and group processing are included in service cost, with report availability defined by Equation C.3. Reducing candidates per call also changes total call count and aggregate output capacity, so results describe a complete grouping policy rather than semantic interference alone.

The relevant tradeoff is whether smaller candidate groups improve accurate detection or rejection enough to compensate for additional sequential calls. We report the same detection/rejection metrics and whole-window service time, rather than substituting single-call latency. Baseline values in Table G.3 use the nested 20-video set and must not be copied from the 100-video tables without re-aggregation.

Table G.3: Candidate grouping on the same nested 20 videos per source, with recent five-frame inputs. Service includes all groups in a window; reports become available at their own group completion.

Model	Condition	Onset F1@3s	Decision F1@3s	Response F1@3s	FAR	Service (s)
CEDVS-Short						
Qwen3-VL-8B	All candidates	45.53	37.40	16.26	20.74	1.153
	Associated groups	34.39	33.12	11.46	42.96	1.856
	One candidate / call	30.06	23.12	4.62	54.81	2.100
InternVL3.5-8B	All candidates	21.54	21.54	6.15	75.56	1.394
	Associated groups	24.31	22.10	7.73	71.11	1.558
	One candidate / call	24.00	24.00	6.86	64.44	2.208
CEDVS-Medium						
Qwen3-VL-8B	All candidates	21.26	24.64	12.08	31.23	0.668
	Associated groups	18.32	20.30	1.49	29.13	1.237
	One candidate / call	11.22	13.63	0.00	45.95	4.846
InternVL3.5-8B	All candidates	8.58	7.40	1.48	76.70	1.126
	Associated groups	8.84	9.70	0.57	79.61	1.667
	One candidate / call	10.58	13.87	0.00	55.99	8.162

Results. Smaller groups have model-dependent benefits and substantial sequential cost. On CEDVS-Medium, InternVL single-candidate input improves onset F1 from 8.58 to 10.58 and reduces FAR from 76.70% to 55.99%, but response F1 falls from 1.48 to zero with 8.162s mean service per window. Qwen instead loses onset F1 (21.26 to 11.22) and increases FAR (31.23% to 45.95%). Associated groups are less expensive than single-candidate calls, yet do not universally improve the joint baseline. CEDVS-Short likewise separates Qwen degradation from modest InternVL changes. The 20-video comparisons describe these fixed subsets and require caution because only 20 videos enter each comparison.

G.5 Uncertainty and Joint Tradeoffs

We perform 2,000 paired video bootstrap replicates with seed 20260922, resampling whole videos with replacement and recomputing micro-averaged onset F1. Table G.4 reports percentage-point differences from the matched five-frame/all-candidate reference and the 2.5th–97.5th percentile interval. Dependencies among events and windows within a video are preserved. Intervals describe sampling uncertainty within the selected subsets, not uncertainty due to run-to-run generation, execution load, annotation error, or model selection; they are not adjusted for multiple comparisons.

Table G.4: Paired onset F1 differences (percentage points) with 95% percentile intervals. Recent-frame/history rows use 100 videos per source; grouping rows use the nested 20 videos.

Condition	CEDVS-Short Qwen	CEDVS-Short InternVL	CEDVS-Medium Qwen	CEDVS-Medium InternVL
Recent 10 frames	-9.9 [-13.5, -6.2]	+1.6 [-0.4, +3.7]	-6.4 [-8.6, -4.2]	+0.4 [-0.4, +1.2]
Recent 20 frames	-20.0 [-23.6, -16.5]	+0.9 [-1.2, +2.9]	-8.9 [-10.7, -7.2]	+0.0 [-0.9, +1.0]
Recent 30 frames	-25.4 [-28.9, -21.8]	-0.9 [-3.0, +1.1]	-11.0 [-12.9, -9.1]	-0.0 [-0.9, +0.9]
History ≤ 32	-25.2 [-28.9, -21.7]	-0.7 [-2.8, +1.3]	-16.2 [-18.7, -14.1]	-2.4 [-3.4, -1.3]
Associated groups	-11.1 [-19.3, -3.8]	+2.8 [-1.5, +7.8]	-2.9 [-7.2, +1.4]	+0.3 [-1.8, +2.9]
One candidate / call	-15.5 [-24.6, -7.5]	+2.5 [-3.0, +8.0]	-10.0 [-14.6, -5.4]	+2.0 [-0.5, +5.1]

What the intervals support. All Qwen recent-frame and history differences exclude zero in the negative direction on both sources. InternVL recent-frame intervals include zero, while its CEDVS-Medium history difference is negative throughout the interval. Although InternVL single-candidate input has a higher point estimate, its interval includes zero on both 20-video subsets. Qwen associated-group input is clearly lower on CEDVS-Short, but its CEDVS-Medium interval includes zero. These distinctions separate consistent losses from tentative gains; none establishes a deployment-optimal grouping policy.

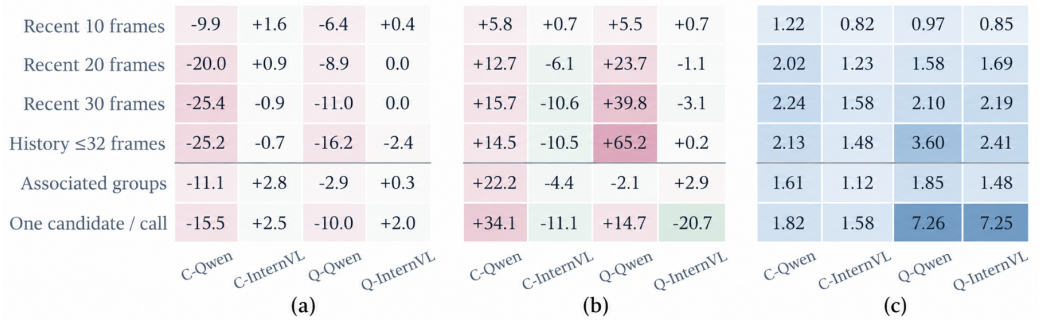

Figure G.1: Joint ablation results relative to the corresponding matched baselines. (a) Change in onset F_1 in percentage points (Δ pp), where positive values indicate improvement. (b) Change in false-alarm rate in percentage points (Δ pp), where negative values indicate improvement. (c) Service-time ratio relative to the matched baseline; values above 1 indicate increased processing cost. C-Qwen and C-InternVL denote results on CEDVS-Short, while Q-Qwen and Q-InternVL denote results on CEDVS-Medium. The first four rows compare different temporal input policies, and the last two rows compare candidate-grouping strategies; the horizontal divider separates these two ablation groups. Service ratios are computed from recorded executions and may reflect differences in system load.

H Representative Successes and Failure Cases

This analysis connects the aggregate metrics to observable model behavior. It distinguishes annotation uncertainty, unavailable visual evidence, semantic decisions, and output/state handling. All cases are selected from saved runs and manually verified. Dataset-quality spot checks in Appendix B do not establish that this model-error review has been performed.

Case selection and evidence. We target four complementary cases: a timely positive trigger with correct rejection of a related absent description; a similar-event false trigger; a retrospectively accurate boundary reported too late; and a missed brief occurrence or failed recovery before a repeated occurrence. Illustrative selection does not estimate population frequencies. Labels may overlap; no error-rate pie chart or category proportion is reported without a separately defined, systematically reviewed sample and denominator.

Figure layout. Each figure places the candidate description above a timestamped sequence of frames, followed by separate ground-truth and prediction timelines. Show START/END transitions and mark the frames actually supplied at the decision under discussion. Future frames included for retrospective explanation must be visually separated and labeled unavailable at that decision. Below the timeline, reproduce a short exact response and annotate $\hat{s}_r$, d_r, and c_r. The completion timestamp is a FIFO simulation using measured service, not a measured live notification. For an

Table H.1: Operational checks for qualitative cases. These are inspection criteria, not measured error counts.

Case type	Evidence required
Timely detection and rejection	A true occurrence, an absent related description, the accepted START, and the other candidate's saved decisions.
Similar-event false trigger	Whole-video absence of the specified event; related visible content and the actual false START.
Late report with accurate onset	True onset s, predicted onset $\hat{s}_r$, decision d_r, and simulated completion c_r, evaluated at the stated tolerance.
Brief or repeated event missed	Ground-truth intervals, actual sampled frames, candidate state, and any preceding END or rejected transition.
Output/interface limitation	Exact response, frozen candidate-ID map, parser audit, and boundary/duplicate handling.
Annotation uncertainty	Visual evidence supporting a disputed description or boundary; record this separately from model failure.

absent candidate there is no ground-truth occurrence interval; show the false report or verified silence without inventing a true onset.

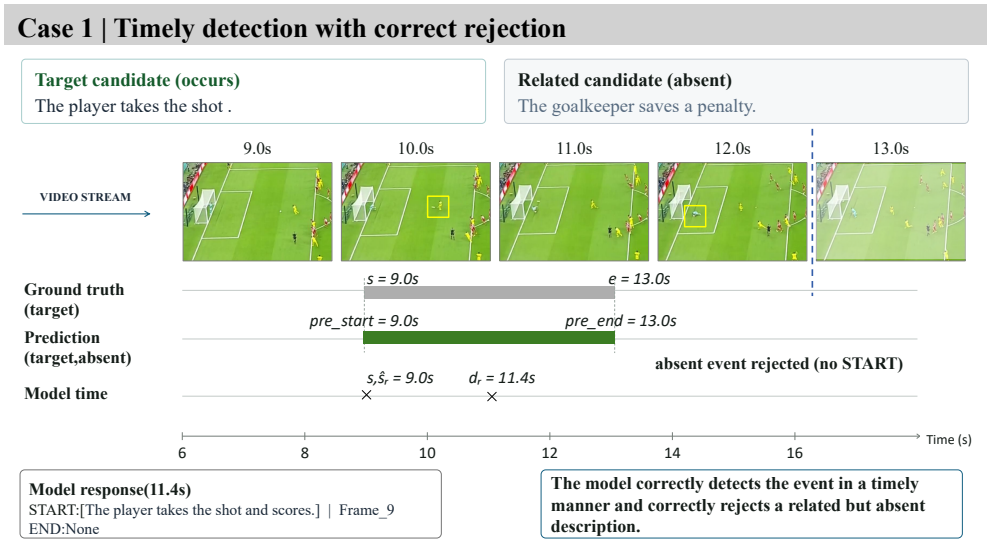

Figure H.1: A qualitative example of timely event detection with correct rejection.

Missing interior samples indicate limited evidence coverage, not proof that inference from context is impossible. Likewise, a missing repeated trigger is attributed to state recovery only when the saved state and transition history support that diagnosis. We describe observable behavior before proposing a mechanism and do not infer a model's internal reasoning from a score alone.

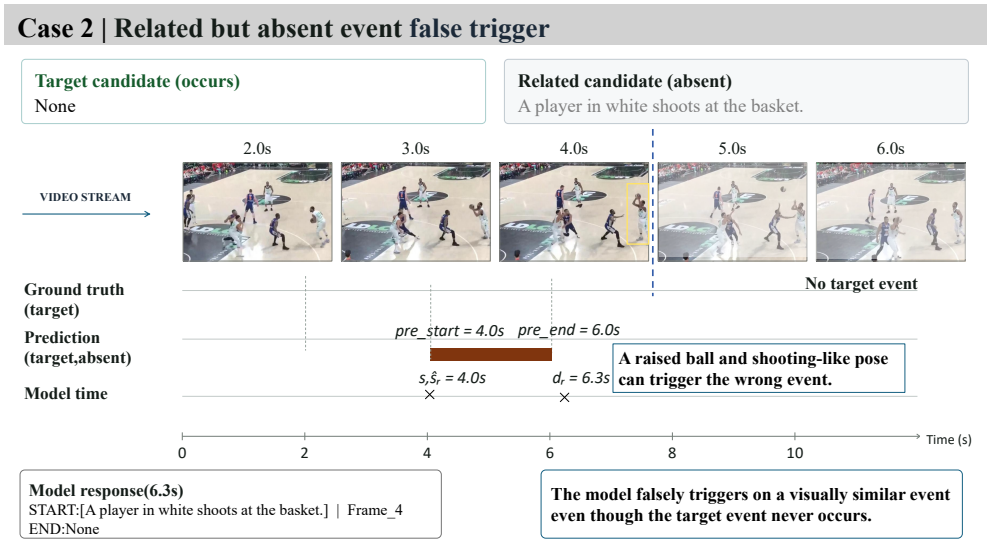

Figure H.2: A qualitative example of a false trigger caused by a visually similar but absent event.

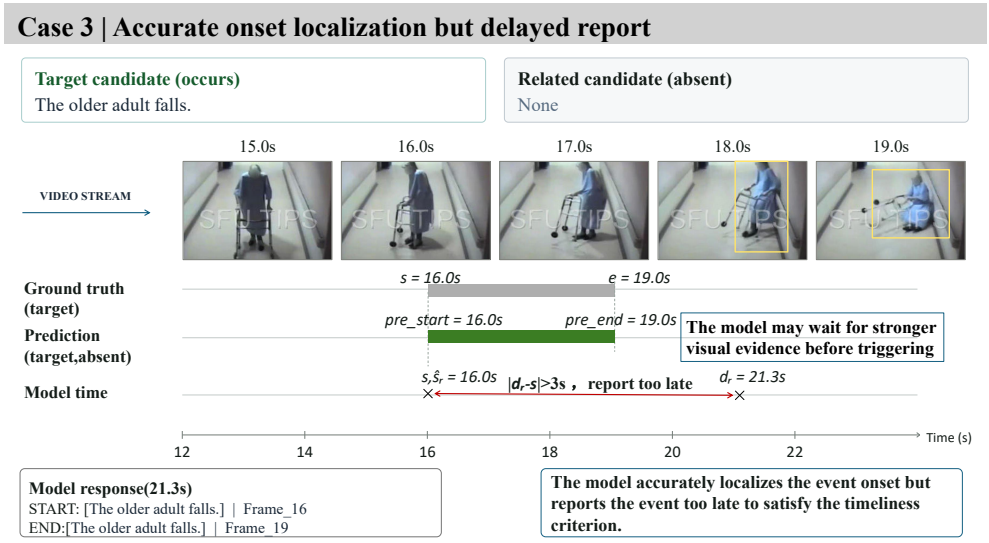

Figure H.3: A qualitative example where the event onset is detected accurately, but the final report is delayed.

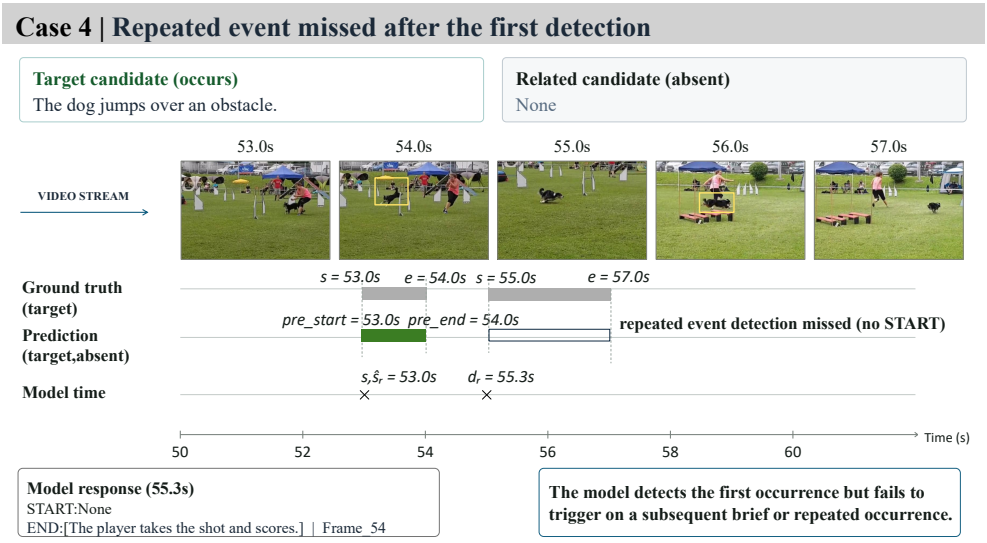

Figure H.4: A qualitative example of a missed brief or repeated event during continuous streaming evaluation.